Computational approaches for understanding one-carbon metabolism in cancer

A Dissertation

Presented to the Faculty of the Graduate School of

Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mahya Mehrmohamadi

January 2017

ABSTRACT

COMPUTATIONAL APPROACHES FOR UNDERSTANDING ONE-CARBON METABOLISM IN CANCER

Mahya Mehrmohamadi, Ph. D.

Cornell University 2017

Cancer metabolism is an emerging research area in cancer biology and therapeutics. One of the major metabolic pathways known to play important roles in the pathogenesis of cancer is one-carbon (1-C) metabolism. 1-C metabolism integrates the status of many dietary nutrients as inputs, and in turn regulates a variety of cellular processes including de novo nucleotide synthesis, lipid metabolism, protein biosynthesis, redox metabolism, transsulfuration, and epigenetics. As the regulation of these cellular processes is critical to cells, the tuning of the activity of 1-C metabolism plays important roles in cancer. Previous studies have established implications of genetic and dietary perturbations of multiple components of 1-C metabolism in human cancers. However, the heterogeneity among cancer types and subtypes with respect to the usage and flux distribution of 1-C metabolism has not been systematically quantified. There remain great potentials in deciphering how 1-C metabolism plays different roles in different human cancers, especially since this metabolic pathway is targeted by a number of the existing antimetabolite chemotherapeutic agents.

In this dissertation, I quantitatively characterize various aspects of 1-C metabolism across human cancers. I first investigate the between-cancer-type variation in the usage of serine by 1-C metabolism using flux distribution analyses and find substantial heterogeneity. I also show that a common feature across cancers is correlated activation of nucleotide and redox metabolism. Next I assess the link between 1-C metabolism and DNA methylation using computational modeling and machine-learning. I find significant contribution from particular enzymes within 1-C metabolism— such as methionine adenosyltransferases— in explaining the within-cancer-type (inter-individual) variation in DNA methylation. My results provide evidence that misregulation of 1-C metabolism is at least in part responsible for disrupted DNA methylation profiles in tumors leading to epigenetic instability and higher malignancy. Given evidence for the role of 1-C metabolism and the methionine cycle in methylation dynamics, I next evaluate the potential for dietary intervention using the amino acid methionine. To this end, I model human serum methionine levels and quantify the contribution of various factors in determining the concentration of methionine. I discover that dietary factors could together explain nearly 30% of overall variation in methionine concentrations, and also provide evidence that the relationship between 1-C metabolism and methylation exists at physiological concentrations of methionine. Finally, I use a novel approach to identify gene expression markers of tumor response to 5-FU and Gemcitabine —two of the commonly used antimetabolite chemotherapies that target enzymes in 1-C metabolism. I discover that response to these agents is to a large degree determined by the metabolic state of tumors and the expression levels of specific target pathways of

each of these agents. Together, my findings provide quantitative information about the heterogeneity among tumors with respect to the usage of 1-C metabolism, and delineate some of the ways this information can be translated into clinical decision-making.

BIOGRAPHICAL SKETCH

Mahya Mehrmohamadi was born in Los Angeles, California on March 1st, 1987. When she was very young, she moved with her family to Iran where they are originally from. Her parents are both professors in academia in Tehran, Iran. When Mahya was in high school, she took a liking for solving problem sets in human genetics. She grew fond of the interdisciplinary nature of this field and how math— which she had been good at since elementary school— can be applied to human health and disease. Also, as a teenager she was very active in community service outside of school. Together with a group of her friends, she worked as a volunteer in a few charity centers including a pediatric cancer hospital in Tehran. Those experiences drew her attention to cancer as a disease that remains difficult and perplexing to clinicians. After finishing high school, like all other high school graduates in Iran, Mahya participated in a national exam for entrance to college. She was ranked 8th nation-wide, which gave her the opportunity to study in any field and at any institution of her choosing in Iran. She chose to pursue her passion for genetics and interdisciplinary science. Mahya got her B.S and M.S in Biotechnology from the University of Tehran in a 5-year combined program. She also worked as a research assistant at Royan stem cell research institute and clinic in Tehran for two years before she decided to continue her education in Genetics and Genomics at Cornell University in 2012. At Cornell, she became very interested in computational biology and decided to switch to more computationally focused research. Mahya joined the lab of Dr. Jason Locasale where she worked on a thesis mainly focused on modeling cancer metabolism using genomics and epigenomics tumor data. She plans to continue

working as a computational cancer scientist and become an independent research investigator in the future.

Dedicated to my grandparents,

Hassan Mehrmohammadi, Mohammad Jafar Pourazar, Sedigheh Giahi, Fatemeh Katouzian

And to my little nephew, Mohammad Safabakhsh

ACKNOWLEDGMENTS

I would first like to express my gratitude to my advisor Dr. Jason Locasale for his continuous support. You have been a great mentor to me, and your advice on both my research and career has been priceless. I owe my progress in scientific writing, research, and communication skills to your patience and guidance. You taught me how to learn from my failures and rise above them, lessons that will help me for the rest of my life. I also want to sincerely thank my co-advisor Prof. Andrew Clark for his encouragement support, not only on matters related to research, but also on other important decisions throughout my PhD. I am especially grateful to you for kindly welcoming me in your lab when the Locasale lab moved to Duke University. It has been an amazing experience and I have learned a great deal from you and the talented members of your diverse research group during this time.

I would also like to thank my thesis committee as well as the faculty at Cornell's MBG department for the great environment they have made that provides students with invaluable resources, helping them develop their careers while in graduate school. I especially like to thank my third committee member Dr. Frank Schroder for all of his support throughout my PhD; and Dr. Jason Mezey for the valuable concepts I learned from his Quantitative Genetics and Genomics course, which I also had the opportunity to TA last year. Finally, I want to thank the faculty at Cornell's Statistics department, especially Drs. James Booth, Martin Wells, and Jacob Bien for useful discussions and collaborations. I also like to thank the team of assistants at Statistical consulting unit and also at Cornell's Cloud Computing center, especially Dr. Brandon Barker for his help with our Red Cloud remote computing account.

I want to thank my all of my collaborators. I am especially grateful to Xiaojing Liu for teaching me experimental metabolomics, Alex Shestov for the flux balance analysis (FBA) work, Lucas Mentch for teaching me about Random Forests, Stephen Salerno for our collaboration on RRmix, and Samantha Mentch for our collaboration on the methionine project. I would also like to thank all current and former Locasale and Clark lab members for their support and useful comments on my research throughout my PhD.

I am especially thankful to my husband Abolhassan Vaezi for his incredible patience and whole-hearted love and support during the past five years. Thank you for always being there for me and reassuring me at times of disappointment and hopelessness! I cannot imagine how I would have been able to survive graduate school without you by my side. You helped me make important decisions when I felt lost and confused, kept me going when I had no idea what I was doing, and helped me keep things in perspective at challenging times. I am truly lucky and grateful for having you.

I would also like to thank my amazing parents Mahmoud Mehrmohamadi and Roya Pourazar for their unconditional love and support. You are truly the best parents I could ask for, and my role models in personal and academic life. Making you proud has always been my biggest motivation for moving toward my dreams. Dad, thank you for always believing in me, even way beyond my potentials! You taught me to aim high, be strong and think big. Mom, thank you for all of the sacrifices you made to keep our home the most peaceful place on Earth! You taught me it is possible to pursue my career goals while building and caring for a family. And special thanks for the amazing packages that you occasionally sent to Ithaca, full of gifts and goodies from home that you and my dear aunt Raheleh prepared for me with love. Seeing

those big yellow Iran post office boxes at our apartment's door always put a big smile on my face!

I also want to thank my family and friends who have contributed appreciably to my happiness and sanity while I was going through grad school. I would first like to thank my sisters. Boshra, thank you for always having helped me see the bigger picture and not get too distracted by the ups and downs of everyday life. Hosna, thank you for always being honest in criticizing me, and special thanks for putting up with all of my craziness while we were roommates for the past year when my husband was away! I should also thank my amazing friend Mona for always being one phone call away and for taking the trouble of traveling to Ithaca to visit me a couple of times. I also want to thank my sweet friend and old neighbor Mina who was truly like a sister to me during the past four years and who made my life in Ithaca absolutely fun and enjoyable.

Finally, I want to thank my funding sources. I was supported by T32GM007617 training grant from the National Institute of Health (NIH), F99 CA212457 grant from the National Cancer Institute (NCI), a grant from International Life Sciences Institute, and a Graduate Fellowship from the Duke University School of Medicine. My research was also supported by grants R00CA168997 and R01CA193256 from the NIH to my advisor Dr. Jason W. Locasale.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION[1]

1.1. Cancer metabolism

Cancer cells alter the usage of their metabolic network by adjusting the flux distribution through the network to accommodate their special needs (Locasale, 2013). This phenomenon is known as "reprogramming" or "rewiring" of metabolism and is one of the recently appreciated hallmarks of cancer (Pavlova and Thompson, 2016) (Hanahan and Weinberg, 2011). Since cancerous cells differ from their normal counterparts with respect to various characteristics, major metabolic shifts are required to equip them with cellular components necessary for proliferation, biosynthesis, stress protection, and survival. The most well-known example of a cancer-specific metabolic program is the famous "Warburg effect" (Vander Heiden et al., 2009) that refers to aerobic glycolysis associated with increased glucose uptake and lactate secretion by cancer cells (Liberti and Locasale, 2016).

Importantly, many metabolic pathways are readily manipulable through enzyme targeting or dietary intervention (Locasale, 2013). Thus, targeting cancer metabolism is an attractive route toward improving cancer therapeutics in more feasible and cost effective ways than most alternative drug development options. To this end, a deeper understanding of metabolic features that distinguish cancers from normal tissues as well as from other cancer types is needed. The increase in the availability of genomics data with the rapid reduction of costs associated with high-

[1] Some of the text in this section has been published in: Mehrmohamadi M, Locasale JW, Context-dependent utilization of serine in cancer. *Molecular and Cellular Oncology*, 2:4, e996418 (2015).

throughput sequencing has made quantitative analyses of tumor metabolism more feasible today than ever before. These valuable technologies facilitate the study of tumor heterogeneity from molecular profiles of tumors.

1.2 One-carbon metabolism in cancer

One-carbon (1-C) metabolism is a metabolic pathway commonly implicated in cancer (Yang and Vousden, 2016) (Locasale, 2013). 1-C metabolism consists of the folate and the methionine cycles that work in concert to integrate nutrient status and availability into cellular metabolism (Figure 1.1). Many dietary factors provide inputs to 1-C metabolism, including folates, amino acids (serine, glycine, and methionine), betaine and vitamin B-12. Thus, the activity of 1-C network is dependent on the levels of these input metabolites, and this network can be thought of as a sensor of the availability of these nutrients (Locasale, 2013). 1-C metabolism in turn regulates the activity of many downstream pathways by producing metabolites that serve as inputs to those downstream pathways (Figure 1.1). Purine and pyrimidine nucleotide synthesis, protein biosynthesis, lipid synthesis, redox metabolism, transsulfuration, and methylation are all critical cellular processes where one or more of the input metabolites are generated through the activity of 1-C metabolism (Locasale, 2013).

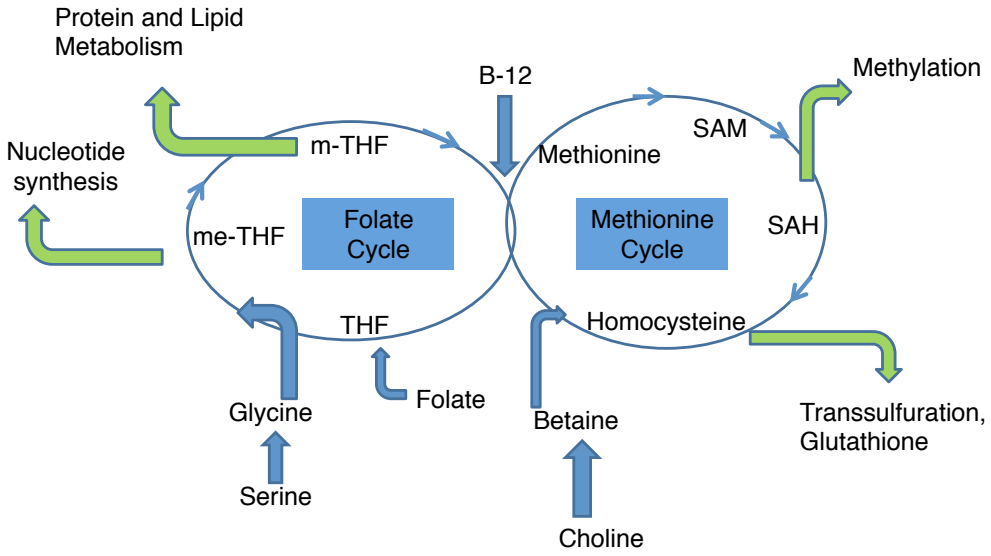


Figure 1. 1— Schematic of the 1-C metabolic pathway.

The Methionine cycle components are shown on the right and the folate cycle on the left. Input metabolites are drawn with blue arrows while the green arrows point to output processes. (Abbreviations: THF= tetrahydrofolate; m-THF= methyl-THF; me-THF= methylene-THF; SAM= S-adenosyl methionine; SAH= S-adenosyl homocysteine).

One-carbon metabolism is highly implicated in cancer (Locasale, 2013). Since 2009 when there was a surge of interest in studying 1-C metabolism in human stem cells as well as cancers (Labuschagne et al., 2014; Locasale et al., 2011; Maddocks et al., 2013; Nilsson et al., 2014; Possemato et al., 2011; Vazquez et al., 2011), numerous studies have provided molecular and epidemiological evidence showing alterations in components of this pathway can play important roles in pathogenesis of multiple human cancers (Locasale, 2013).

1.2.1 Serine metabolism

Serine and glycine donate single carbon units and fuel the 1-C metabolic network. Through the action of the 1-C network, these nutrients support a wide variety of downstream cellular processes such as nucleotide synthesis, methylation metabolism, sulfur metabolism, polyamine metabolism, lipid and protein synthesis and redox balance (Locasale, 2013). In recent years, cancer researchers have reported multiple instances where serine uptake or its de novo synthesis was significantly elevated in tumor cells (Jain et al., 2012; Maddocks et al., 2013) (Locasale et al., 2011; Possemato et al., 2011). It was shown that several human cancer types including breast and colorectal cancers rely on serine for proliferation and survival and that hyperactivity in the network can drive the development of cancer. The utilization of serine for increasing de novo nucleotide synthesis rates in highly proliferative cells has been the main proposed explanation for this observation. However, serine and glycine feed into numerous other metabolic pathways as well. Recent studies have shown that serine is also used for production of NADPH and that the role of serine in the regulation of the redox status could be very critical to cell proliferation (Fan et al., 2014; Lewis et al., 2014; Nilsson et al., 2014; Ye et al., 2014). Intermediates in 1-C metabolism including folate, betaine, and cystathionine in the serum have all been linked to proliferation of cancer cells (Locasale, 2013). Together, evidence supports the importance of serine metabolism in many instances of human cancers (Yang and Vousden, 2016).

1.2.2 The methionine cycle

Methylation is a key biochemical reaction and many molecules in cells undergo this modification. Misregulated methylation of proteins, nucleic acids, and metabolites contribute to many human conditions (Bergman and Cedar, 2013; Greer and Shi, 2012; Kraus et al., 2014). When methylation occurs at histones and DNA that determine the epigenetic status in cells, it can affect gene expression programs (Barth and Imhof, 2010). Previous studies provide evidence that compared to normal counterparts, cancer cells exhibit elevated disordered variation in their DNA methylation patterns (Timp and Feinberg, 2013). This suggests a model of cancer as a dysregulated epigenome that provides tumor cells with epigenetic plasticity, acting in parallel with genetic instability to assure higher adaptive potential in cancer compared to normal cells. Methylation levels can be altered due to dysregulated expression or activity of methyltransferase and demethylase enzymes in cancer (Chi et al., 2010; Dawson and Kouzarides, 2012). It has also been long established that S-adenosylmethionine (SAM) is the universal methyl substrate for these enzymes that transfer its methyl group to yield a methylated product, and is itself converted to S-adenosylhomocysteine (SAH) (Finkelstein, 1990). This provides a link between SAM production and the epigenetic status of cells (Gut and Verdin, 2013; Katada et al., 2012; Teperino et al., 2010). A study from our research group provided direct evidence for the importance of the methionine cycle in regulating histone methylation in both healthy and cancerous tissues (Mentch et al., 2015). Together, previous studies have illustrated an important regulatory link between 1-C metabolism and cellular epigenetics suggesting great potential for metabolic manipulation of cellular gene expression dynamics.

1.2.3 Chemotherapies that target 1-C metabolism

Many widely used chemotherapeutic agents that are approved by the U.S. Food and Drug Administration (FDA), usually classified as cytotoxic and non-specific, in fact target specific enzymes within 1-C metabolism (Locasale, 2013). A few of the well-known examples include the anti-folate Methotrexate which targets dihydrofolate reductase (DHFR), and nucleotide analogs 5-fluorouracil (5-FU) and Gemcitabine which interfere with nucleotide biosynthesis though inhibiting 1-C metabolic enzymes thymidylate synthetase (TYMS) and ribonucleotide reductase (RRM), respectively (Amelio et al., 2014). These agents can often be tolerated and can achieve remarkable responses in advanced stage cancers leading to complete remission in many cases. However, the clinical responses to these agents are heterogeneous with some patients exhibiting complete resistance. Thus, a better characterization of 1-C metabolism has great potential for improving and personalizing the administration of currently existing antimetabolite drugs. Results from previous analyses of the power of 1-C metabolic genes in predicting patient response to antimetabolites are largely controversial (Etienne-Grimaldi et al., 2010; Vazquez et al., 2013; Zhao et al., 2016). Despite numerous studies on cancer cell lines as well as cancer patients, the clinical administration of antimetabolite agents remains mainly non-specific (Audet-Walsh et al., 2016; Iorio et al., 2016; Ser et al., 2016). No metabolic biomarkers are currently used in practice for stratifying patients based on whether or not they are likely to benefit from these agents. With the increased availability of molecular data on human

tumors, there is considerable potential in computational assessment of tumor genomic data in search for determinants of response to antimetabolite chemotherapies.

1.3 Metabolic heterogeneity in cancer

Numerous molecular studies have illustrated roles for components of 1-C metabolism in cancer (Amelio et al., 2014), however, the heterogeneity among different human cancer types and subtypes with respect to the usage and reprogramming of 1-C metabolism remains largely uncharacterized. It is well established that cancer is a complex disease and populations of tumor cells show substantial levels of both inter- and intra-tumor variation that makes cancer treatment highly challenging (McGranahan and Swanton, 2015). In recent years, many studies have quantified aspects of tumor heterogeneity by capitalizing on the availability of genome-scale information on tumors leading to interesting clinical findings (Andor et al., 2016). Knowledge provided by studies of this kind can guide clinicians in making more informed decisions in cancer diagnostics, choice of therapy regimens, personalizing chemotherapies, and predicting cancer outcome. Similar to other molecular features of cancer, 1-C metabolism is also highly variable among different cancer types and subtypes (Hu et al., 2013). The heterogeneity of cancer cells with respect to 1-C metabolism has not been systematically characterized and quantified to date. Important questions that remain to be further investigated include heterogeneity in the usage of different metabolic inputs, the contribution from 1-C metabolism in regulating methylation dynamics, and flux distribution through 1-C metabolism in different human cancers.

1.4 In this dissertation

In this work, I focus on characterizing the variability among human cancers with respect to 1-C metabolism. More specifically, I quantify differences between individual tumors with respect to flux distribution, amino acid utilization, epigenetic regulation, drug response, and patient survival based on 1-C activity in tumors.

In chapter two, I use gene expression profiles of hundreds of human tumors to predict serine flux distribution through 1-C metabolism and find considerable heterogeneity in the usage of serine across human cancers. Notably, I find that nucleotide synthesis and redox metabolism are co-regulated in cancers. I then assess the validity of these expression-based flux predictions using experimental serine tracing and metabolomics in cancer cell lines and confirm that fluxes can be predicted from pathway-level gene expression data.

In chapter three, I study the link between 1-C metabolism and epigenetic regulation in cancer. I build computational models of DNA methylation using hundreds of molecular features across thousands of human tumors. I then quantify the contribution of 1-C metabolism in explaining variation in DNA methylation at multiple levels and find a surprisingly large contribution for 1-C metabolic genes—especially for enzymes in the methionine cycle. Finally, I assess the clinical implications of these results by performing survival analysis and find that tumors in which DNA methylation is regulated by 1-C metabolism tend to be less malignant than tumors in which this link appears to be disrupted.

In chapter four, I study the variability of the amino-acid methionine —which provides the chemical link between 1-C metabolism and epigenetics— in the human serum using metabolomics and computational modeling. This study finds substantial variability among individuals in their serum methionine levels that could potentially impact methylation and epigenetics. Results confirm the validity of dietary interventions for manipulating methionine levels in the human serum. Finally, using predictive modeling and analysis of variance, I quantify the clinical, dietary, and physiological determinants of methionine levels in the serum.

Lastly in chapter five, I investigate the extent of specificity in the action of antimetabolite chemotherapies that target 1-C metabolism, and introduce a framework for identifying gene expression markers of patient response to a number of these agents. I show that this approach is more powerful than individual gene expression markers in identifying responder and non-responder sub-groups of patients. I also find sets of metabolic pathways that proved valuable in stratifying patient subgroups with respect to response to 5-FU and Gemcitabine. Together, my results illustrate that antimetabolite agents act more specifically than previously appreciated, and there is great unexplored potential in further searching for biomarkers within the specific metabolic pathways that these agents target, using novel approaches such as the one applied to gene expressions in the current work.

Overall, this dissertation aims at systematic and quantitative analyses of major roles of 1-C metabolism in cancer, with a focus on heterogeneity among and within cancer types. Results provide valuable insights into clinical applicability of uncovering

such variability in cancer, and also introduce computational strategies for tackling such complex questions using high-dimensional genomic data.

1.5. References

Amelio, I., Cutruzzola, F., Antonov, A., Agostini, M., and Melino, G. (2014). Serine and glycine metabolism in cancer. Trends in biochemical sciences *39*, 191-198.

Andor, N., Graham, T.A., Jansen, M., Xia, L.C., Aktipis, C.A., Petritsch, C., Ji, H.P., and Maley, C.C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature medicine *22*, 105-113.

Audet-Walsh, E., Papadopoli, D.J., Gravel, S.P., Yee, T., Bridon, G., Caron, M., Bourque, G., Giguere, V., and St-Pierre, J. (2016). The PGC-1alpha/ERRalpha Axis Represses One-Carbon Metabolism and Promotes Sensitivity to Anti-folate Therapy in Breast Cancer. Cell reports *14*, 920-931.

Barth, T.K., and Imhof, A. (2010). Fast signals and slow marks: the dynamics of histone modifications. Trends in biochemical sciences *35*, 618-626.

Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. Nature structural & molecular biology *20*, 274-281.

Chi, P., Allis, C.D., and Wang, G.G. (2010). Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. Nature reviews. Cancer *10*, 457-469.

Dawson, M.A., and Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell *150*, 12-27.

Etienne-Grimaldi, M.C., Milano, G., Maindrault-Goebel, F., Chibaudel, B., Formento, J.L., Francoual, M., Lledo, G., Andre, T., Mabro, M., Mineur, L., et al. (2010). Methylenetetrahydrofolate reductase (MTHFR) gene polymorphisms and FOLFOX response in colorectal cancer patients. British journal of clinical pharmacology *69*, 58-66.

Finkelstein, J.D. (1990). Methionine metabolism in mammals. The Journal of nutritional biochemistry *1*, 228-237.

Greer, E.L., and Shi, Y. (2012). Histone methylation: a dynamic mark in health, disease and inheritance. Nature reviews. Genetics *13*, 343-357.

Gut, P., and Verdin, E. (2013). The nexus of chromatin regulation and intermediary metabolism. Nature *502*, 489-498.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., Vander Heiden, M.G., and Vitkup, D. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nature biotechnology *31*, 522-529.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Goncalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740-754.

Katada, S., Imhof, A., and Sassone-Corsi, P. (2012). Connecting threads: epigenetics and metabolism. Cell *148*, 24-28.

Kraus, D., Yang, Q., Kong, D., Banks, A.S., Zhang, L., Rodgers, J.T., Pirinen, E., Pulinilkunnil, T.C., Gong, F., Wang, Y.C., et al. (2014). Nicotinamide N-methyltransferase knockdown protects against diet-induced obesity. Nature *508*, 258-262.

Labuschagne, C.F., van den Broek, N.J., Mackay, G.M., Vousden, K.H., and Maddocks, O.D. (2014). Serine, but Not Glycine, Supports One-Carbon Metabolism and Proliferation of Cancer Cells. Cell reports *7*, 1248–1258.

Liberti, M.V., and Locasale, J.W. (2016). The Warburg Effect: How Does it Benefit Cancer Cells? Trends in biochemical sciences *41*, 211-218.

Locasale, J.W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. Nature reviews. Cancer *13*, 572-583.

Locasale, J.W., Grassian, A.R., Melman, T., Lyssiotis, C.A., Mattaini, K.R., Bass, A.J., Heffron, G., Metallo, C.M., Muranen, T., Sharfi, H., et al. (2011). Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. Nature genetics *43*, 869-874.
Maddocks, O.D., Berkers, C.R., Mason, S.M., Zheng, L., Blyth, K., Gottlieb, E., and Vousden, K.H. (2013). Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. Nature *493*, 542-546.
McGranahan, N., and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer cell *27*, 15-26.
Mentch, S.J., Mehrmohamadi, M., Huang, L., Liu, X., Gupta, D., Mattocks, D., Gomez Padilla, P., Ables, G., Bamman, M.M., Thalacker-Mercer, A.E., et al. (2015). Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. Cell metabolism *22*, 861-873.
Nilsson, R., Jain, M., Madhusudhan, N., Sheppard, N.G., Strittmatter, L., Kampf, C., Huang, J., Asplund, A., and Mootha, V.K. (2014). Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. Nature communications *5*, 3128.
Pavlova, N.N., and Thompson, C.B. (2016). The Emerging Hallmarks of Cancer Metabolism. Cell metabolism *23*, 27-47.
Possemato, R., Marks, K.M., Shaul, Y.D., Pacold, M.E., Kim, D., Birsoy, K., Sethumadhavan, S., Woo, H.K., Jang, H.G., Jha, A.K., et al. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature *476*, 346-350.
Ser, Z., Gao, X., Johnson, C., Mehrmohamadi, M., Liu, X., Li, S., and Locasale, J.W. (2016). Targeting One Carbon Metabolism with an Antimetabolite Disrupts Pyrimidine Homeostasis and Induces Nucleotide Overflow. Cell reports *15*, 2367-2376.
Teperino, R., Schoonjans, K., and Auwerx, J. (2010). Histone methyl transferases and demethylases; can they link metabolism and transcription? Cell metabolism *12*, 321-327.
Timp, W., and Feinberg, A.P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nature reviews. Cancer *13*, 497-510.
Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science *324*, 1029-1033.
Vazquez, A., Markert, E.K., and Oltvai, Z.N. (2011). Serine biosynthesis with one carbon catabolism and the glycine cleavage system represents a novel pathway for ATP generation. PloS one *6*, e25881.
Vazquez, A., Tedeschi, P.M., and Bertino, J.R. (2013). Overexpression of the mitochondrial folate and glycine-serine pathway: a new determinant of methotrexate selectivity in tumors. Cancer research *73*, 478-482.
Yang, M., and Vousden, K.H. (2016). Serine and one-carbon metabolism in cancer. Nature reviews. Cancer *16*, 650-662.
Zhao, T., Xu, Z., Gu, D., Wu, P., Huo, X., Wei, X., Tang, Y., Gong, W., He, M.L., and Chen, J. (2016). The effects of genomic polymorphisms in one-carbon metabolism pathways on survival of gastric cancer patients received fluorouracil-based adjuvant therapy. Scientific reports *6*, 28019.

CHAPTER 2: CHARACTERIZATION OF THE USAGE OF SERINE IN HUMAN CANCER[2]

2.1 Abstract

The serine, glycine, one carbon (SGOC) metabolic network is implicated in cancer pathogenesis but its general functions are unknown. I carried out a computational reconstruction of the SGOC network and then characterized its expression across thousands of cancer tissues. Pathways including methylation and redox metabolism exhibited heterogeneous expression indicating a strong context dependency of their usage in tumors. From an analysis of coexpression, simultaneous up- or down-regulation of nucleotide synthesis, NADPH and glutathione synthesis was found to be a common occurrence in all cancers. Finally, my collaborators and I developed a method to trace the metabolic fate of serine using stable isotopes, high-resolution mass spectrometry and a mathematical model. Although the expression of single genes did not appear indicative of flux, the collective expression of several genes in a given pathway allowed for successful flux prediction. Together these findings identify expansive and heterogeneous functions for the SGOC metabolic network in human cancer.

[2] Mehrmohamadi, M., Liu, X., Shestov, A.A. & Locasale, J.W. Characterization of the usage of the serine metabolic network in human cancer. *Cell Reports* **9**, 1507-1519 (2014).

2.2 Introduction

Serine and glycine are nutrients that fuel metabolic pathways including one carbon metabolism and sulfur metabolism. This metabolic unit referred to as the serine, glycine and one carbon (SGOC) network provides an integration point in cellular metabolism that allows for cells to achieve diverse biological functions by converting serine and glycine into several metabolic outputs. These outputs include building blocks for nucleotide, lipid, and protein synthesis. They also include polyamine synthesis and work in the maintenance of redox status as determined by glutathione biosynthesis and NADPH production (Circu and Aw, 2010; Fan et al., 2014; Lewis et al., 2014a, b; Murphy et al., 2011; Tedeschi et al., 2013). The network also provides the substrates for methylation reactions that may have relevance to maintaining cellular epigenetic status (Gut and Verdin, 2013; Teperino et al., 2010).

Recent work has pointed to new roles of the SGOC network in cancer pathogenesis (Chaneton et al., 2012; Jain et al., 2012; Ma et al., 2013; Maddocks et al., 2013; Scuoppo et al., 2012; Zhang et al., 2012). Whereas a subset of cancer cells increase de-novo serine biosynthesis (Locasale et al., 2011; Possemato et al., 2011), other cancers benefit from an increased serine and glycine uptake rate which allows them to metabolize these amino acids for their biosynthetic needs (Jain et al., 2012; Maddocks et al., 2013). Importantly, a recent study showed that serine but not glycine is critical in providing the one-carbon units required for biosynthesis of nucleotides in some cancer cells (Labuschagne et al., 2014). A critical role for the mitochondrial folate pathway in rapidly proliferating cancer cells has also been recently elucidated

(Nilsson et al., 2014). Furthermore, the importance of one-carbon metabolism in NADPH production through the oxidation of folates was demonstrated in cancer cells in a number of recent studies (Fan et al., 2014; Lewis et al., 2014a, b; Tedeschi et al., 2013; Vazquez et al., 2011). These studies showed that in addition to providing cells with nucleotide units, one-carbon metabolism has an important role in redox balance. Despite these advances, the general coordinated usages and different contexts in which serine and glycine flux contributes to different metabolic functions within and across cancer types and normal tissues remain largely unknown.

Previous studies have analyzed the expression levels of metabolic pathways across a variety of cancer types using meta-analysis approaches (Hu et al., 2013; Nilsson et al., 2014; Tedeschi et al., 2013; Vazquez et al., 2013). These studies have identified tumor-associated changes in gene expression across human metabolism including one carbon metabolism. Whereas previous work has characterized expression of the SGOC network effectively as a pathway (Hu et al., 2013) and thus one or two data points or as a single series of individual genes (Nilsson et al., 2014), I attempted to carefully examine with higher resolution the coordination and context-dependence of functionally distinct pathways of serine utilization. I further investigate my computational findings experimentally by tracing the metabolic fate of serine and connecting these observations to expression patterns in the network. Together I identify several novel context dependent utilizations of serine in human cancers.

2.3 Results

2.3.1 Reconstruction of the human SGOC network

I constructed a network that represents the metabolism of serine and glycine through one carbon metabolism and other immediate pathways including the transsulfuration pathway that together lead to defined cellular outputs. This network, collectively referred to the SGOC network, was generated first by curating all human metabolic genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). Each gene involved in the KEGG-defined pathways, Glycine - Serine and Threonine metabolism, Cysteine and Methionine metabolism, and Folate biosynthesis was then selected. I subsequently included genes involved in adjacent chemical reactions (edges) of the selected genes (nodes), and these nodes were then connected to the selected edges to allow for a contiguous sequence of chemical reactions. Finally, I pruned this network to exclude each isolated node as well as enzymes that carry out chemical reactions involved in other pathways that I concluded to be non-specific to SGOC metabolism (Figure 2.1A, Methods).

As a result, the network comprises of a set of sixty-four genes involving core metabolic reactions and enzymes and isoenzymes (Figure 2.1B). Inputs of the network include de novo serine and glycine metabolism from glucose and the import of serine and glycine from the extracellular space. Outputs of the network include purine, pyrimidine, lipid, glutathione, redox, taurine, and methylation metabolism. In practice the network is compartmentalized into cytosolic/nuclear and mitochondrial components that can be to some extent captured in these reaction sequences when metabolites are shuttled in and out of the mitochondria.

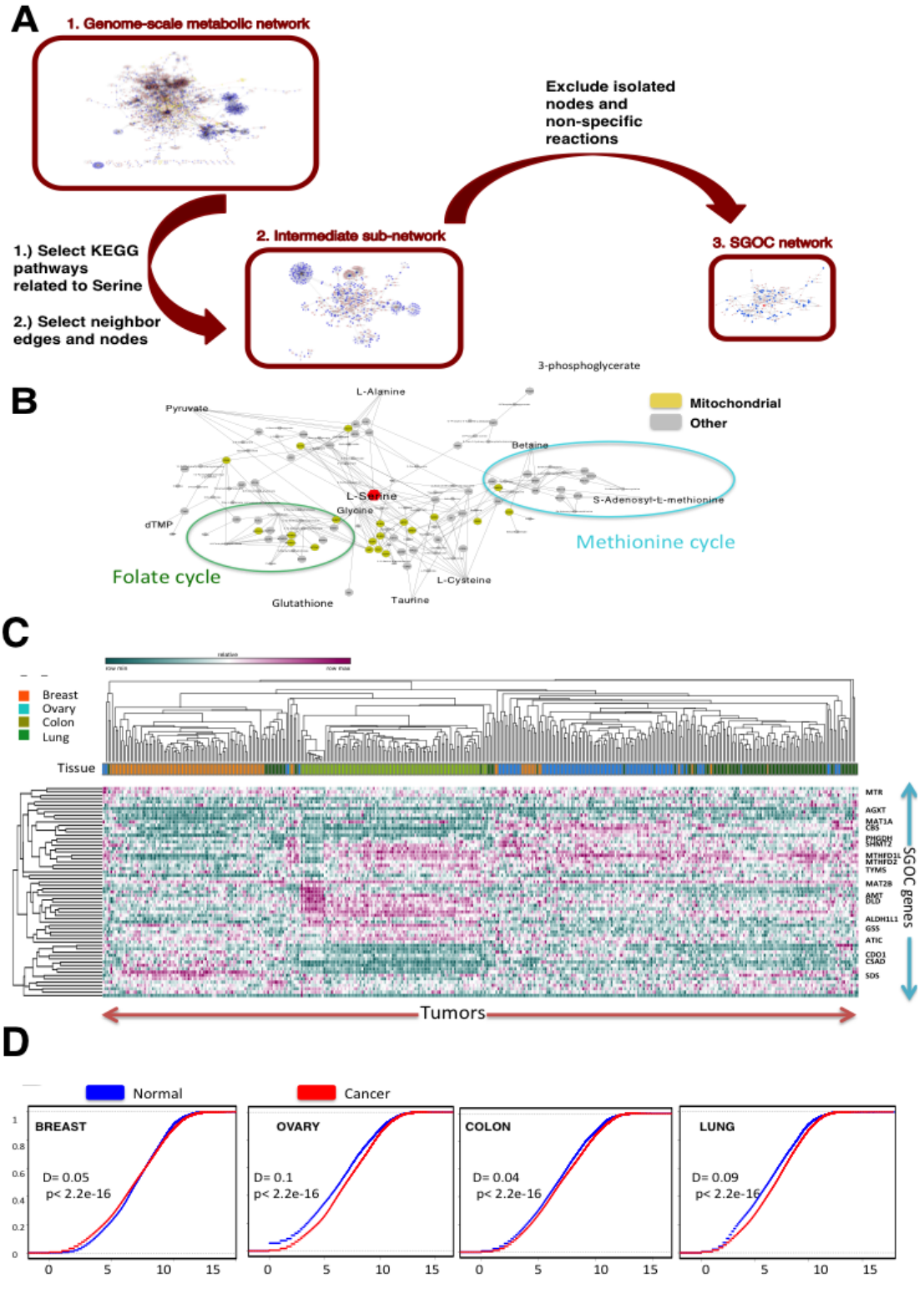

A
1. Genome-scale metabolic network
Exclude isolated nodes and non-specific reactions
1.) Select KEGG pathways related to Serine
2.) Select neighbor edges and nodes
2. Intermediate sub-network
3. SGOC network
B
3-phosphoglycerate
L-Alanine
Pyruvate
Mitochondrial
Other
Betaine
L-Serine
Glycine
S-Adenosyl-L-methionine
dTMP
Methionine cycle
Folate cycle
L-Cysteine
Glutathione
Taurine
C
Breast
Ovary
Colon
Lung
Tissue
MTR
AGXT
MAT1A
CBS
PHGDH
SHMT2
MTHFD1L
MTHFD2
TYMS
MAT2B
AMT
DLD
ALDH1L1
GSS
ATIC
CDO1
CSAD
SDS
SGOC genes
Tumors
D
Normal
Cancer
BREAST
D= 0.05
p< 2.2e-16
OVARY
D= 0.1
p< 2.2e-16
COLON
D= 0.04
p< 2.2e-16
LUNG
D= 0.09
p< 2.2e-16

Figure 2.1 – Reconstruction of the functional SGOC network.

A) The entirety of the human metabolic network is first considered using a list of all metabolic genes from KEGG. A sub network involving all human serine, glycine, and one carbon metabolism related genes are considered obtained from KEGG. Reaction paths from this network were used to create the final SGOC network.

B) Schematic of resulting network and its decomposition into functional outputs (mitochondrial isoforms are shown in yellow).

C) Expression patterns of SGOC network genes across four cancer types (ovarian, lung, colorectal, and breast) visualized with unsupervised hierarchical clustering (100 randomly chosen samples from each cancer type from TCGA were used). (See also Figure A1.S1).

D) Cumulative density plots showing the distribution of SGOC expression levels across four tumor types and their corresponding normal tissues (GENT).

2.3.2 Expression of the SGOC network in human cancers and normal tissues

Having constructed this network, I then investigated its expression in several cancer types and corresponding normal tissues. I considered both The Cancer Genome Atlas (TCGA) data across Breast, Ovarian, Lung, and Colorectal cancers (Cancer Genome Atlas, 2012a, b; Cancer Genome Atlas Research, 2011, 2012) and the Gene Expression in Normal and Tumor (GENT) database (Shin et al., 2011) that contains larger numbers of normal tissue samples. I chose these cancer types for two reasons: 1) The three major types Breast, Lung, and Colon represent major sources of cancer mortality with ovarian cancer exhibiting expression patterns similar to that of subtypes of breast cancer observed to have enhanced requirements for serine metabolism (Locasale et al., 2011; Possemato et al., 2011) and 2) Each of the three major cancer types considered has been shown to exhibit clinical response to chemotherapies that target nodes within the one carbon metabolism network (Chabner

and Longo, 2011). For each tumor type, expression data are available across hundreds of tumors allowing for extensive statistical characterization.

I first assessed the global expression of the genes in the network (Table 2.1). A hierarchical clustering across the cancer types revealed clustering based largely on tissue type (Figure 2.1C) consistent with the results of principal component analysis (PCA) (Figure A1.S1). Major exceptions were observed for lung cancer samples as well as a subset of breast cancers that clustered with ovarian cancer in which case breast cancers lacking estrogen receptor exhibited expression patterns indistinguishable from those of ovarian cancer. Next, an analysis of the global properties of the network was considered. In ovarian, colon, and lung cancers, the overall distribution of SGOC network gene expression is shifted toward higher expression levels compared to the levels in corresponding normal tissue (Figure 2.1D). However, in breast cancer, the broad range of the cumulative density plot shows a higher variability in the expression levels of SGOC genes compared to normal breast tissue (Figure 2.1D). In all cancer types, the distribution of SGOC expression differed between tumor and normal (Kolmogorov-Smirnov p-value < 2.2e-16). I next considered the variability between and within cancer types to further identify cancer contexts of the network (Table 2.1). These contexts include high expression in one tumor type relative to others, large variation in a single tumor type, high expression in tumor versus corresponding normal tissue, and high variability in tumor vs. normal. The resulting calculations revealed that genes predominantly involved in de novo nucleotide synthesis pathways are over-expressed in all four cancer types compared to the corresponding normal tissues (Table 2.1). This result is in agreement with the

previous literature emphasizing the importance of nucleotide biosynthesis to rapidly proliferating cells (Hu et al., 2012; Wilson et al., 2012). Furthermore, consistent with a recent observation (Nilsson et al., 2014) I also observed methylenetetrahydrofolate dehydrogenase 2 (NADP+ dependent) (MTHFD2) and serine hydroxymethyltransferase 2 (SHMT2), two mitochondrial enzymes that contribute to nucleotide metabolism, to be consistently overexpressed in cancers compared to corresponding normal tissues (Table 2.1). Furthermore, when SGOC enzymes with a mitochondrial isoform were compared to their cytosolic isoforms, the mitochondrial isoforms showed stronger up-regulation in tumors (Fisher's exact test p-value= 0.02), demonstrating the importance of mitochondrial compartment in tumor metabolism (Supplementary Methods). However, some cancer type-specific expression patterns were also observed. For instance, cystathionine beta-synthase (CBS), serine dehydratase (SDS), and glutathione synthetase (GSS) were expressed more highly in ovarian, breast, and colon tumors, respectively (Table 2.1). These three genes also showed very little within-cancer type variation in contrast to other genes. Cross normal tissue comparisons revealed high expression of thymidylate synthetase (TYMS) in normal colon and alanine-glyoxylate aminotransferase (AGXT) in normal breast (Table 2.1). Together, the variation in genes within tumor types and in comparison to corresponding normal tissues suggested many newly identified context dependent expression patterns in the network.

Table 2.1. SGOC expression analysis across tissues.

Name	T-T highes	T-T significa	N-N highes	N-N significa	T-N overexp.	T-N significa

	t	nce	t	nce	(pvalue<2.8e10-6)	nce
ADC	breast	x+	breast	x+	breast,ovary,lung	-,-,-
AGXT1	colon	x+	breast	x+	Lung	-
AGXT2(m)	breast	x+	breast	x+	colon,lung	-,-
AHCY	colon	x	colon	x+	breast, colon, lung	-,+,+
AHCYL1	ovary	x+	breast	x+	ovary,colon	-,-
ALAS1(m)	colon	x	colon	x+	Ovary	-
ALAS2(m)	lung	x	breast	x+	Colon	-
AMT(m)	colon	x	ovary	x	-	
ANPEP	colon	+	breast	x+	Ovary	-
ATIC	colon	x	colon	x+	breast, colon, lung	+,+,+
BHMT	lung	x	breast	x+	ovary,lung	-,-
BHMT2	breast	x+	breast	x+	ovary,lung	-,-
CBS	ovary	x+	ovary		breast, ovary,lung	-,-,+
CDO1	breast	x+	breast	x+	Colon	-
CHDH(m)	colon	x	breast	x+	ovary,colon, lung	-,-,+
CP	ovary	x	lung	x	ovary,lung	-,+
CSAD	breast	x+	breast	x+	-	
CTH	colon	x+	colon	x+	lung	+
DHFR	colon	x	colon	x+	breast,ovary,colon,lung	-,-,+,+
DLD(m)	colon	x+	colon	x+	breast, ovary	-.-
DMGDH(m)	ovary	x+	breast	x+	colon	-
FPGS(m)	ovary	+	breast	x+	ovary,colon,lung	-,+, -
FTCD	ovary	x	breast	x+	ovary,colon,lung	-,-,-

GAD1	lung	x	breast	x+	colon, lung	+,+
GAD2	lung		breast	x+	colon,lung	-,-
GAMT	breast	x	breast	x+	ovary,lung	-,+
GART	colon	x	breast	x	breast,ovary,colon ,lung	-,+,+,+
GATM	breast	x+	breast	x+	lung	-
GCAT(m)	lung		colon	x	breast,colon,lung	-,-,+
GGH	colon	x	colon	x+	breast,ovary,colon ,lung	-,-,-,+
GLDC(m)	ovary	x+	breast	x+	ovary,lung	-,+
GNMT	breast	x	breast		-	
GOT1	colon	x+	colon	x+	breast	-
GOT2(m)	colon	x+	ovary	x	breast, colon, lung	-,-,-
GPT	colon	x+	colon	x+	-	
GPT2	ovary	x	breast	x+	colon, lung	+,+
GSS	colon	x+	colon	x+	breast, colon, lung	-,-,+
IL4I1	ovary	x	ovary		breast,ovary,colon ,lung	-,-,-,+
MAT1A	ovary	x+	ovary	x	colon	-
MAT2A	ovary	x+	lung		ovary,colon	+,-
MAT2B	colon	x	breast	x	-	
MPST(m)	colon	x	colon	x+	ovary	-
MTFMT (m)	colon	x	ovary		-	
MTHFD1	colon	x+	breast	x+	colon,lung	-,+
MTHFD1L (m)	lung	x	ovary	x	colon, lung	+,+

MTHFD2 (m)	lung	x	colon	x	breast,ovary,colon ,lung	+,+,+,+
MTHFR	ovary	x+	breast	x	-	
MTHFS	lung	x	ovary		breast, colon, lung	-,-,-
MTR	ovary	x+	ovary	x+	colon	+
NAGS(m)	colon	x	colon		lung	-
PDPR(m)	ovary	x+	breast	x+	ovary,colon	-,-
PHGDH	ovary	x	breast	x+	colon,lung	+,-
PIPOX	ovary	x	ovary	x+	colon, lung	+,+
PPCS	lung		ovary	x	breast,lung	-,+
PPIG	ovary	x+	ovary	x	-	
PSAT1	ovary	x	breast	x+	ovary,colon,lung	+,+,+
PSPH	lung		ovary	x	breast,colon	-,+
SARDH(m)	breast	x+	breast	x+	lung	-
SDS	breast	x+	breast	x+	breast,ovary,colon ,lung	-,-,-,+
SHMT1	ovary	x	breast	x+	lung	-
SHMT2(m)	ovary	x+	ovary	x	breast,ovary, colon, lung	+,-,+,+
TYMS	colon		colon	x+	breast,colon, lung	+,-,+
ALDH1L1	colon	x	breast	x+	ovary,colon,lung	-,+,+
ALDH1L2 (m)	NA	NA	breast	x+	ovary,colon,lung	+,+,+

The SGOC genes are listed in the first column with mitochondrial enzymes denoted by "m". The second column shows the cancer type with the highest average expression across the 4 cancers in the study. In the significance column, Mann-Whitney-Wilcoxon p-values smaller than the genome-level Bonferroni threshold (2.8 e 10-6) are denoted by an "x" and effect sizes larger than 0.3 are denoted by a "+" sign. The fourth and fifth columns summarize similar analysis across the 4 normal tissue types. The sixth and seventh columns summarize the results of tumor vs. normal comparisons. In the sixth column, tissue types in which a

significant over-expression (p-value < 2.8 e 10-6) was detected in tumors are listed. The fifth column specifies corresponding effect sizes by a "+" if larger than 0.3 and a "-" otherwise. "NA" denotes missing data.

2.3.3 Serine utilization in the SGOC network

After analyzing the expression of individual genes, I considered a functional analysis of the metabolic outputs of the network. I first introduced a framework for understanding the different metabolic fates of serine. I decomposed the SGOC network into pathways that utilize serine as an input and achieve a distinct biological function as an output (Methods). I also added de novo serine biosynthesis as an additional pathway due to its implications in cancer (Locasale, 2013). For each pathway, I analyzed the overall range of expression (Figure 2.2A). It was observed that there is large within-cancer variability in expression of pathways especially across different breast tumors (Figure 2.2A). Furthermore, mean pathway expression was similar across cancer types for some pathways such as methylation, whereas there were significant differences in other pathways including de novo serine biosynthesis (Figure 2.2A). Next I developed an algorithm to evaluate the expression along each functional pathway (Figure 2.2B). I first decomposed each biologically distinct unit of the network into a set of genes. Then, the mean, median, and variation of expression for each gene comprising each pathway was computed and statistics were evaluated across the population of tumors (Methods). Since the pathways varied in length, I normalized the obtained values to the number of genes contained in the pathway.

I considered this pathway analysis with several contrasts: expression levels in one cancer relative to other tumor types (T-T), variability in individual tumor (T) types, over-expression of the pathway in tumor relative to normal tissue (T-N), expression of the pathway in one normal tissue relative to other tissue types (N-N), variability in tumor vs. corresponding normal tissue (T-N CV). Each of these analyses provides unique information relevant to cancer. For example, high expression in the normal tissue may indicate a predisposition for the usage of the pathway in a particular cancer type. Furthermore, a high variability indicates possible selective usage or overexpression in some context of tumorigenesis such as a particular mutational event. Due to the tremendous heterogeneity within cancer types, a high within-cancer-type variability is important to know since it implies that different sub-populations of samples behave differentially with respect to a certain pathway, which suggests their potential use as biomarkers, provides further context for the use of the pathway, and finally possibly yields some predictive capacity for evaluating the response to agents that target the pathway. A comparison across tumor tissues shows that several pathways are overexpressed relative to other tumors with breast and ovarian cancer exhibiting higher expression of specific components of sulfur related metabolism (Figure 2.2C). High variability across each cancer type was observed throughout the network but the taurine, methylation, and NADPH pathways showed highest variability compared to other pathways (Figure 2.2D). However, when comparing expression in normal and tumor tissue, only routes related to nucleotide and redox metabolism were commonly upregulated in tumors consistent with previous analysis of individual genes (Figure 2.2E). Interestingly, in comparing

individual genes and pathways, I found that high relative levels of expression in one normal tissue did not necessarily overlap with high relative levels of expression in tumor tissue, indicating shifts in metabolism that are not directly due to differences in tissues of origin (Figure 2.2F). This observation is also apparent in comparing the variability in expression of tumor and normal cancer types (Figure 2.2G). Together these findings identify novel relationships of context dependent utilizations of the SGOC network.

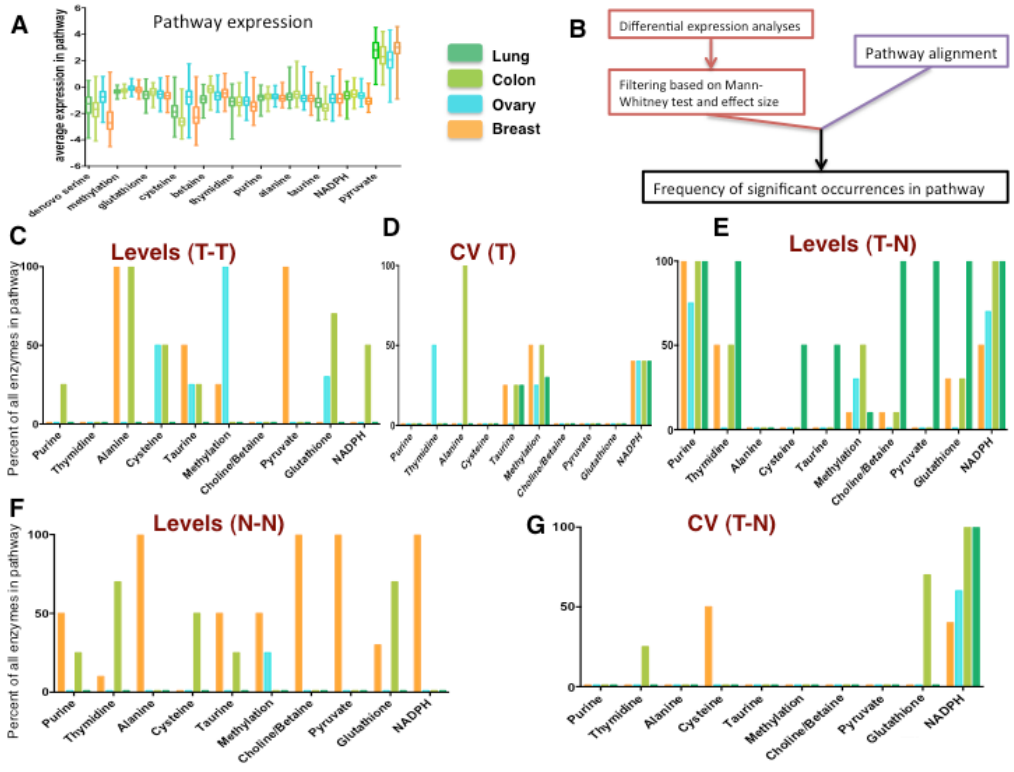


Figure 2.2 – Functional SGOC pathway utilization across four cancer types.

A) Boxplots showing average pathway expression across four cancer types (TCGA).

B) Summary for quantification of pathway utilization analyses.

C-G) Bar-plots denoting the collective expression of a given pathway route in the SGOC network. C) Pathways hyper-utilized in a single cancer type relative to others. D) Pathways exhibiting high variation within cancer types. E) Pathways overexpressed in tumor versus corresponding normal tissue. F) Pathways highly expressed in one normal tissue relative to others. G) Pathways with high variation in tumor versus normal tissue.

2.3.4 Co-occurrence of pathway utilization within the network

Since in metabolism, the output of one branch of the network is coupled to the output of all other branches, I hypothesized that there could exist correlations in the expression of sets of genes leading to metabolic outputs of the network. To investigate this possibility I computed a similarity matrix for each of the cancer types. Clustering of each similarity matrix revealed co-occurring expression patterns in each cancer type suggesting coordinated utilization of certain enzymes (Figure 2.3A). A global analysis of these pairwise correlations revealed that the genes in the SGOC network were significantly more correlated compared to randomly chosen genes as revealed by quantile-quantile (QQ) plots (Figure 2.3B). To my knowledge this is the first systematic demonstration of coordinated expression of metabolic genes in a defined metabolic pathway in humans.

Next, I considered the normalized mean expression of the pathway routes that illustrated the extent of co-occurrence in the network (Figure 2.3C). When assessed across the four cancer types, it was found that correlations emerge along specific pathways. Surprisingly many of these correlations were largely found to be independent of tissue type demonstrating the existence unique metabolic programs within each tissue. A core group of pathways involving glutathione, NADPH, and

nucleotide metabolism show correlations in their expression suggesting that a coupling between de novo nucleotide and redox metabolism occurs. De novo serine synthesis correlated only with cysteine metabolism in some cancer types indicating that the condensation of serine and entry into the transsulfuration pathway is a major usage of serine in cells with enhanced de novo serine biosynthesis. Also in all cancer types, taurine and nucleotide metabolism anti-correlated with one another indicating that their usages are orthogonal in cancers.

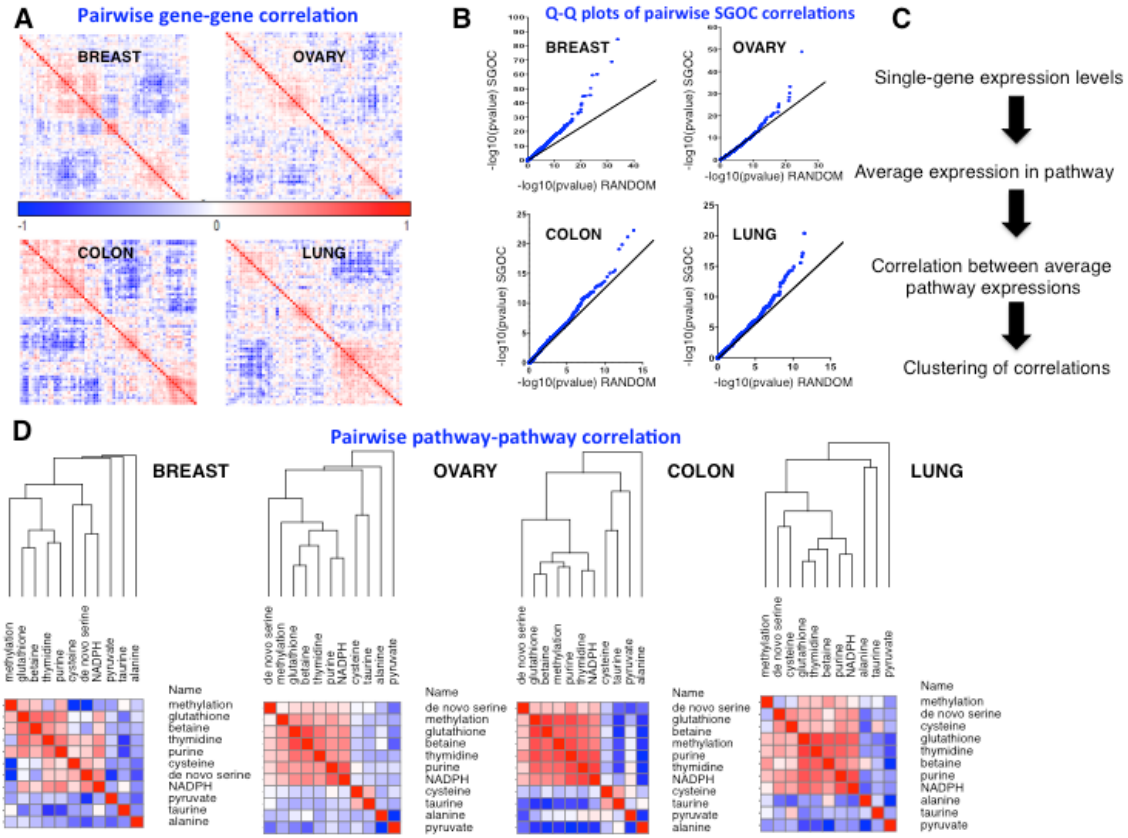


Figure 2. 3 – Correlations across reaction paths in the SGOC network.

A) Hierarchically-clustered similarity matrices for four cancer types based on pairwise Spearman correlations. Modularity in the co-expression network is apparent.

B) Overview of algorithm that collapses co-regulated genes into reaction paths.

C) Quantile-Quantile (QQ) plots showing p-values from gene-gene Spearman correlations in SGOC compared to randomly chosen pair of genes.

D) Correlations in expression of transcripts in one pathway in reference to another. Results are organized by hierarchical clustering using Spearman correlations in linkage similarities.

2.3.5 Serine-derived metabolic fluxes in the network

Gene expression in human clinical samples offers an unbiased assessment of the expression of the network in cancer. However, gene expression does not necessarily determine metabolic phenotype that ultimately involves metabolic flux or the rate of flow of a metabolite from one point in the network to another. To investigate the relationship of metabolic flux with gene expression, my collaborator Xiaojing Liu first developed a ^{13}C-serine-based isotopomer mass spectrometry method. We incubated a panel of cancer cells with ^{13}C labeled serine and measured relative abundances of all mass isotopomers (molecules differing only in the extent of their isotopic composition) of each metabolite from the integrated ion current. Uniformly labeled serine (U-^{13}C Serine) was present in the media in a 1:1 proportion with respect to unlabeled serine in the culture medium. Therefore, at steady state, as expected a substantial portion of the serine is unlabeled (Figure A1.S2). Furthermore, M+1 glycine was detected near the natural abundance level in my experiment (Figure A1.S2), suggesting that glycine cleavage does not happen in reverse at a substantial rate. The eight human colon cancer cell lines used in this study consistently showed

labeling of glutathione and de novo purine and thymidine synthesis intermediates, while little or no label was detected on methionine, pyruvate, alanine, betaine and taurine (Figure 2.4A, Figure A1.S3). Since gene expression is available for each of these cell lines, I could then ask to what extent patterns in gene expression could be related to these isotope patterns. Using abundance ratios of mass isotopomers as surrogates for the corresponding fluxes, I studied correlations between labeling patterns and gene expression first at the single gene level and then at the level of pathways (Figure 2.4B). I asked whether gene expression within the network could predict isotope-labeling patterns. I found that the expression of individual genes in the SGOC network could to some extent predict fluxes (Figure 2.4C, FDR q-value 0.15). Notably, results from a Fisher's exact test indicate that gene expression at the pathway level is more strongly correlated with fluxes compared to single gene level (Figure 2.4D). In summary, results from these data indicate that fluxes can be determined directly to an extent from gene expression in cancer cells but this requires knowledge of the overall expression of the pathway. To my knowledge this is the first study in a mammalian system that provides an example of the relationship between gene expression and flux.

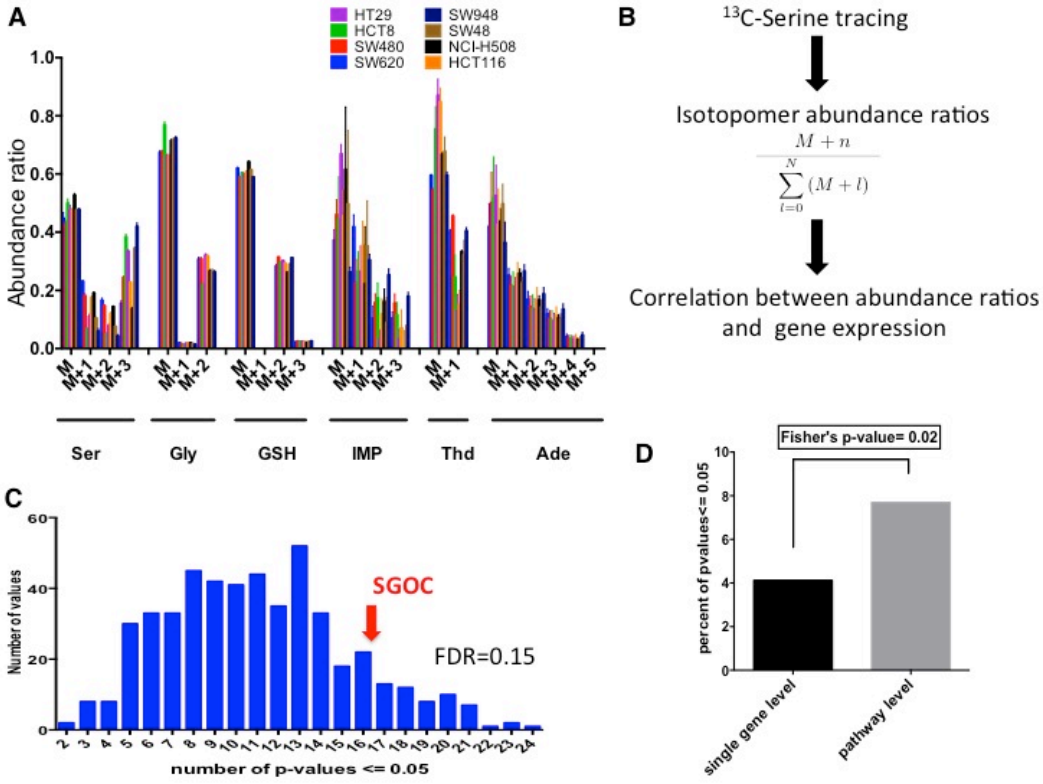


Figure 2.4 – Flux analysis of serine metabolism.

A) Abundance ratios for labeled isotopomers detected across the 8 colon cancer cell lines.

B) Overview of mass isotopomer analyses.

C) Histogram of the number of significant correlations between labeled SGOC metabolite isotopomers and expression of randomly selected networks of genes. Expression of SGOC genes shows a higher than random correlation (FDR=0.15).

D) Quantification of the frequency of significant correlations between SGOC expression and isotopomer labeling at single genes level vs. pathway level using a Fisher's exact test. (See also Figure A1.S3).

2.3.6 Interaction between de novo nucleotide and glutathione biosynthesis

Next, I further investigated the relationship between de novo nucleotide metabolism and glutathione synthesis (Figure 2.5A) in colorectal cancer cells. I considered the relationship between thymidine and glutathione synthesis, purine and glutathione synthesis, and purine and thymidine synthesis (Figure 2.5A). In each case, the expression levels of key genes in the pathway provided little information about the pathway relationships. However, when comparing pathway expression, strong correlations in nucleotide synthesis and glutathione biosynthesis exist. This phenomenon was observed both in TCGA colon cancer data as well as in gene expression data from the 8 colon cancer cell lines in the study. Experimentally, this result is also apparent in that labeling of de novo nucleotide synthesis intermediates was highly correlated with glutathione labeling (Figure 2.5A). Together these findings point to a model where increases in nucleotide biosynthesis are coupled to flux to glutathione synthesis whereas other pathway routes from serine are not correlated with nucleotide synthesis (Figure 2.5B). Since NADPH synthesis is coupled to the redox balance of oxidized and reduced glutathione, this finding provides a possible connection to a recent finding indicating that one carbon metabolism is a major source of NADPH in cells (Fan et al., 2014). In fact, I also observe a positive association between NADPH synthesis with nucleotide and glutathione synthesis at the pathway level (Figure A1.S4).

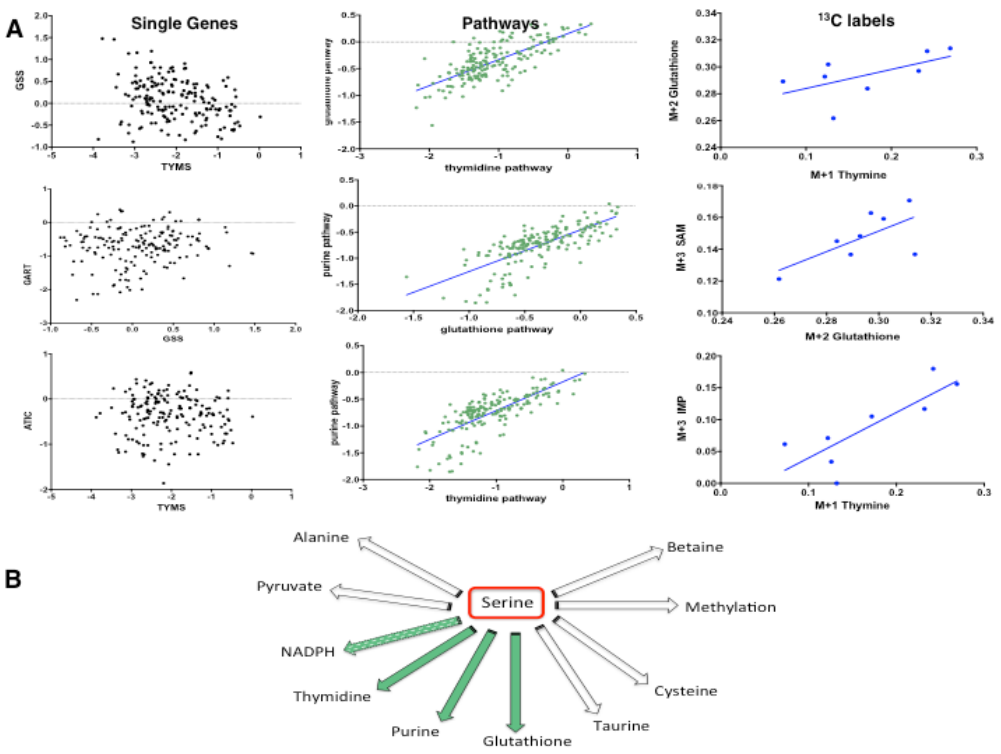


Figure 2.5 – Interaction between de novo nucleotide and glutathione synthesis.

A) Thymidine and Glutathione (top panel): The left plot shows weak association between TYMS and GSS expression (R-squared=0.094). The middle plot shows significantly higher correlation between thymidine and glutathione at average pathway expression level (R-squared= 0.53). The right plot shows association between glutathione and thymine label ratios (R-squared=0.35). Purine and Glutathione (middle panel): The left plot shows weak association between phosphoribosylglycinamide formyltransferase (GART) and GSS expression at single gene level (R-squared=0.01). The middle plot shows significantly higher correlation between purine and glutathione at average pathway expression level (R-squared=0.54). The right plot shows association between glutathione and S-adenosyl-methionine (SAM) - a purine intermediate- label ratios (R-squared=0.44). Thymidine and Purine (bottom panel): The left plot shows weak association between TYMS and IMP cyclohydrolase (ATIC) expression at single gene level (R-squared=0.001). The middle plot shows significantly higher correlation between thymidine and purine at average pathway expression level (R-squared=0.55). The right plot shows association between inosine-monophosphate (IMP) -a purine intermediate- and thymine label ratios (R-squared=0.67).

B) Schematic showing serine metabolic outputs. Results from tracing serine in vitro suggest a simultaneous flux going through de novo nucleotide and glutathione synthesis pathways (green arrows) in the colon cancer cell lines studied when little or no flux goes from serine to the other pathways. NADPH labeling could not be assayed in my experimental set-up using ^{13}C-serine.

2.3.7 Mathematical modeling of pathway fluxes

Although labeling patterns can sometimes be used to infer the extent of flux through a pathway, in general this is not always the case. Fluxes ultimately are required to be estimated from network models that account for the isotope patterns. I therefore considered a mathematical model of the Mass Isotopomer Distributions (MIDs) to estimate quantitative values for the fluxes in the network (Figure 2.6A, Methods). A two-compartment Serine-Glycine-Methylenetetrahydrofolate (Ser-Gly-meTHF) metabolic model was built by my collaborator Alexander Shestov, to fit experimental ^{13}C MIDs of the SGOC network to determine metabolic fluxes involving transport and exchange relative to extracellular serine transport (Ftr-ser=1). The metabolic network includes cellular serine production via de novo synthesis from 3-phosphoglycerate, extracellular serine uptake, reversible cytosolic SHMT1 and mitochondrial SHMT2 fluxes, mitochondrial glycine cleavage system activity (GCS), serine, glycine and meTHF exchange fluxes between cytosolic and mitochondrial compartments, dilution fluxes for cytosolic and mitochondrial one-carbon metabolite pool represented in the model by meTHF, de novo and salvage formation pathways for adenine (Ade representing purines) and thymidine (Thd representing pyrimidines), and glutathione production flux. An analysis of the fluxes across each cell line provided direct quantitative information about the distribution of fluxes across the network (Figure 2.6B). For example, flux from glycine to serine varied over an order of magnitude suggesting that some cells may be able to compensate for differences in

serine availability through the uptake of glycine. Across the cell lines, results from the analyses of estimated fluxes show a positive relationship between the de novo purine and thymidine pathway fluxes consistent with the pathway expression levels across the 8 cell lines (Figure 2.6C). Furthermore, using estimated flux through SHMT1 as a surrogate for glutathione production flux, a positive relationship is detected between glutathione and thymidine synthesis both at the flux level as well as the pathway expression level (Figure 2.6D). Similarly, purine biosynthesis shows a positive association with glutathione production (Figure 2.6E). Together these findings both identify the heterogeneity of pathway fluxes across the network in a set of cell lines and provide confirmation of correlations that exist in the utilization of serine for specific metabolic functions.

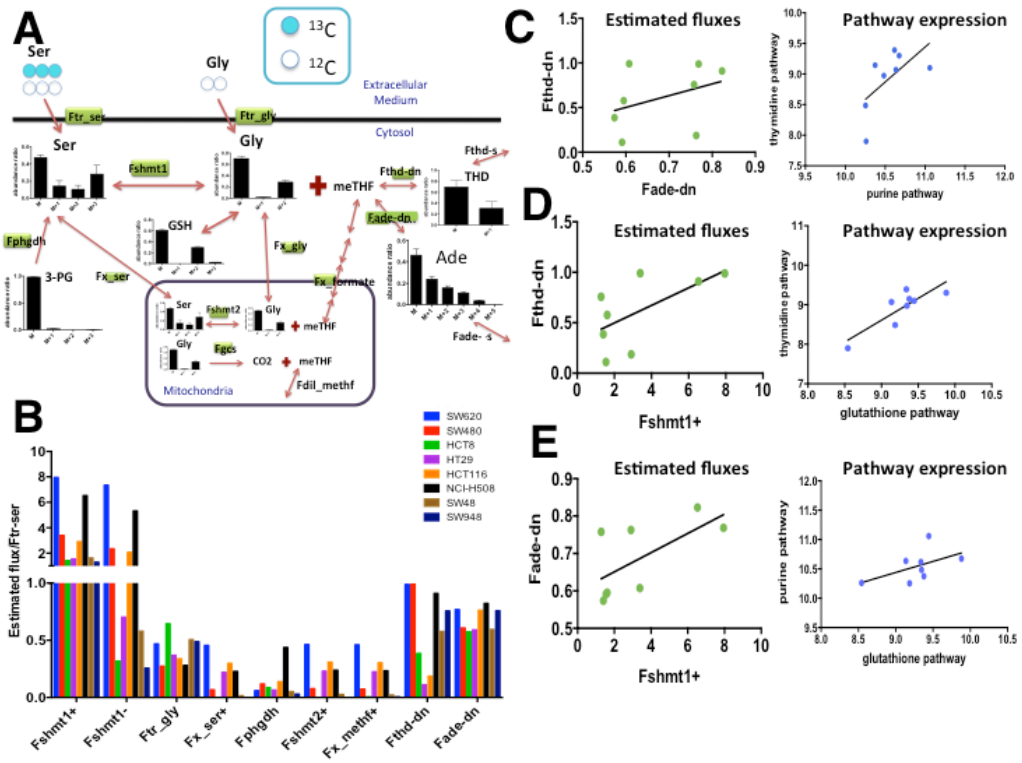


Figure 2.6 – Mathematical modeling estimates fluxes in purine and pyrimidine synthesis pathways.

A) Schematic of the model that was used for flux estimation. Serine, glycine, and formate are shuttled in and out of the mitochondria whereas folate is not. Plots show mass isotopomer distributions (MID) for metabolites that were detected experimentally. Fluxes that were estimated are labeled by green rectangles.

B) Barplots of estimated fluxes with respect to the serine transport flux (Ftr-ser=1) across the 8 cell lines. (Fshmt1+: forward flux through SHMT1; Fshmt1-: reverse flux through SHMT1; Ftr-gly: Glycine transport flux; Fx-ser+: serine exchange flux; Fphgdh: de novo serine synthesis flux; Fshmt2+: forward flux through SHMT2; Fx-methf+: meTHF exchange flux through formate; Fthd-dn: de novo thymidine synthesis flux; Fade-dn: de novo adenine synthesis flux)

C) Purine and thymidine pathways are correlated. Estimated fluxes for de novo thymidine and adenine synthesis are positively associated (R-squared=0.14) (left). Expression of the purine pathway is positively correlated with that of the de novo thymidine pathway (R-squared=0.4) across the 8 colon cancer cell lines (right).

D) Glutathione and thymidine pathways are correlated. Estimated fluxes for de novo thymidine and glutathione synthesis (SHMT1+) are positively associated (R-squared=0.39) (left). Expression of the glutathione pathway is positively correlated with that of the de novo thymidine pathway (R-squared=0.7) across the 8 colon cancer cell lines (right).

E) Glutathione and purine pathway correlations. Estimated fluxes for de novo adenine and glutathione synthesis (SHMT1+) are positively associated (R-squared=0.42) (left). Expression of the glutathione pathway is positively correlated with that of the de novo purine pathway (R-squared=0.28) across the 8 colon cancer cell lines (right).

2.4 Discussion

My findings present the first comprehensive systems-level analysis of the expressions patterns, metabolic fluxes, and interrelationships of serine metabolizing pathways in numerous cancer contexts and together delineate the likely roles of the SGOC network in several cancer types. While there is no general global overexpression of the entire network in cancer as found in other pathways such as glycolysis, the expression of the network differs in more complex ways. Heterogeneity is strikingly apparent with for example breast cancers showing both

overall under- and over- expression of multiple pathway components. Components contributing to nucleotide synthesis are commonly upregulated as previously reported (Hu et al., 2013; Jain et al., 2012), but other pathways exhibit no general features of over- and under- expression across cancer types. In normal tissues, many of these components are constituently expressed with variability suggesting that the usage of a particular component in the normal tissue may confer predisposition of its usage in cancer.

When considering co-expression of the network, strong correlations in individual genes and expression of the pathways were observed. The most common co-existence was the expression of genes contributing to nucleotide metabolism, glutathione and NADPH synthesis. This was also observed in direct measurements of pathway fluxes for glutathione and nucleotides. Since glutathione is the essential component of cellular redox maintenance, concomitant synthesis of glutathione and nucleotides likely provides a redox environment necessary for nucleotide synthesis and repair. In fact, several studies in plants have demonstrated that redox regulation by glutathione has an important role during nucleotide synthesis and cell division (Belmonte et al., 2003; Belmonte et al., 2005; Stasolla, 2010). By performing co-expression analyses on the serine pathways, I provided new insights into the biology of the SGOC network in cancer. In fact, my finding that glutathione synthesis is associated with nucleotide synthesis appears novel in the context of human cancer, and could suggest that cancer cells utilize a redox balance mechanism in parallel to the up-regulation of biosynthetic pathways.

Gene expression measurements are typically thought to not be useful surrogates in cells for flux measurements that are the ultimate phenotypes of interest in studying metabolism. However, I show that pathway-level gene expression to a large extent in the SGOC network can be used as a predictor of experimentally measured fluxes. This is a very important finding in general given the wealth of publically available gene expression data, and the technical limitations associated with performing large-scale metabolomics analyses on human tumor samples. Therefore this work hopefully provides motivation for further comparative analysis of gene expression and flux distributions in biological samples.

2.5 Methods

2.5.1 Cell culture and metabolite extraction

Colon cancer cells (SW620, SW480, HCT8, HT29, HCT116, NCI-H508, SW48, and SW948) were cultured as previously described (Liu et al., 2014). For ^{13}C-serine tracing experiments, cells were seeded in 6-well plate at a density of 2×10^5 to 5×10^5 cells per well. After overnight incubation in 37 °C with 5% CO_2, full growth media were removed, and cells were washed with 2 ml PBS before the addition of RPMI media supplemented with 10% dialyzed and heat inactivated FBS, 100 U/ml penicillin, 100 mg/ml streptomycin and ^{13}C-U-serine (3 mg/100 ml medium, Cambridge Isotope Laboratory) such that in the final medium, 50% of serine is ^{13}C labeled. After a 24 hour incubation, the cells were harvested as previously described (Liu et al., 2014).

2.5.2 Mass spectrometry and Liquid chromatography

The qExactive Mass Spectrometer (QE-MS) coupled to liquid chromatography (Ultimate 3000 UHPLC) was used for metabolite separation and detection as previously described (Liu et al., 2014). Raw data collected from LC-Q Exactive MS was processed with Sieve 2.0 (Thermo Scientific). Relevant instrument parameters are contained in a previous work (Liu et al., 2014).

2.5.3 Network construction and gene expression analyses

The network reconstruction was carried out using the Cytoscape plugin – Metscape (Gao et al., 2010). The SGOC network was generated first by curating all human metabolic genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). All genes involved in the KEGG-defined pathways Glycine, Serine and Threonine metabolism, Cysteine and Methionine metabolism, and Folate biosynthesis were selected. I subsequently included genes involved in adjacent chemical reactions (edges) of the selected genes (nodes), and these nodes were then connected to the selected edges to allow for a contiguous sequence of chemical reactions. Finally, I manually excluded all genes that involve non pathway-specific chemical reactions that thus take part in many metabolic pathways implying that their function need not be confined to 1-C metabolism (reactions including: ATP-dependent; tRNA synthesis; Methylation; Co-A; CO2 and NH3; aldehyde dehydrogenases; pyruvate metabolism) (see Table S2 for the complete list of genes excluded).

Gene expression across human tumors was analyzed using level 3 TCGA (The Cancer Genome Atlas) mRNA data (Cancer Genome Atlas, 2012a, b; Cancer Genome Atlas Research, 2011, 2012). Data from AgilentG4502A microarray chips were collected for all TCGA breast (BRCA), ovarian (OV), colon (COAD), and lung (LUSC, LUAD) samples corresponding to four cancer types. Level 3 data contain combined probe signals for each gene and samples were LOESS normalized using a reference RNA sample (cy5/cy3)(Yang et al., 2002), therefore, poor probe binding was accounted for by reporting the ratio mRNA from sample to that of the standard.

Mean and median expression levels were calculated for each of the 64 genes in the SGOC network across each cancer type.

For evaluating high expression of one cancer type relative to other cancer types, I considered the following. For each gene, the cancer type that had the highest median expression was grouped ("high expression") and all samples from the other three cancer types were pooled into one group ("low expression"). The Mann-Whitney Wilcoxon (Wilcoxon) test was then performed to compare gene expression between the "high expression" vs. "low expression" groups. The Bonferroni method for multiple hypothesis correction was then applied to determine the significant p-values. Effect sizes for the Wilcoxon test are calculated as r=Z/sqrt(N) where N is the total number of samples and Z is the Z-score for the Wilcoxon test. Genes with a significant p-value but a small effect size ($r < 0.3$) were also considered insignificant. The result of this analysis is a set of genes highly expressed in one cancer type.

For tumor vs. normal comparisons, Affymetrix U133Plus2 expression data from the GENT (Gene Expression across Normal and Tumor tissues) dataset were used for breast, ovarian, colon, and lung cancer as well as normal tissue samples. The data were preprocessed using the MAS5 algorithm and then normalized to a target density of 500 to correct for batch effects according to GENT specifications (Shin et al., 2011). Each cancer type was compared to its corresponding normal tissue using the same statistical approach (Wilcoxon test with Bonferroni and effect size corrections). Comparisons of expression in one normal tissue type relative to other normal tissue types were conducted on these data using the same statistical criterion.

For an analysis of variability in expression, the coefficient of variation (CV) was calculated for each gene across each tissue type. CV comparisons for each case were carried out using the same criteria starting with the Wilcoxon test.

2.5.4 Pathway definitions and analysis

I considered the amino acid serine as the input source of one-carbon units to the SGOC network and decomposed the network into biochemical pathways that consume serine in different reactions and result in a distinct biological output. The enzymes in each of the pathways from serine metabolism are as follows (mitochondrial isoforms are denoted by "m"):

Methylation: SHMT1/SHMT2 (m) – MTHFR- MTR/BHMT/BHMT2-MAT1A/MAT2A/MAT2B

Thymidine: SHMT1/SHMT2(m) – AMT(m) -TYMS- DHFR

Purine: SHMT1/SHMT2(m) - MTHFD1/MTHFD2(m) /MTHFD1L(m) - ATIC/GART

NADPH: SHMT1/SHMT2(m) - MTHFD1/MTHFD2(m) /MTHFD1L(m) - ALDH1L1/ALDH1L2(m)

Alanine: AGXT1/AGXT2(m)

Glutathione: SHMT1/SHMT2(m) - GSS

Cysteine: CBS- CTH

Taurine: CBS- CTH- CDO1- CSAD

Betaine/Choline: SHMT1/SHMT2(m) - CHDH(m)

Pyruvate: SDS

To compare these pathways in reference to the statistical analysis that I conducted for each gene, I considered an overall pathway score (frequency of occurrences). For each pathway, the number of significant genes, normalized to the number of genes in that pathway, was computed for each of the statistical analyses that I performed. Results are reported in Figure 2.2.

For the analysis of network covariance, similarity matrices were constructed based on pairwise Spearman correlations between expression levels of the 64 SGOC metabolic genes across each one of the four cancer types (Lung, Breast, Colon, and Ovarian) in the study. In order to visualize these correlations in the context of serine-fate pathways, average pathway expression levels were measured for all tumor samples in study, and similarity matrices were made based on Spearman correlations between average pathway expressions separately in each cancer type. All clustering calculations were carried out using the Gene-E package (Broad Institute www.broadinstitute.org/cancer/software/GENE-E/).

For comparison of pairwise correlations between expression levels of SGOC genes to that expected by chance, I used a randomization method. I randomly picked 64 genes from the genome and calculated pairwise correlations of those genes and repeated this for 100 iterations. Finally, I used the average of sorted p-values from the

100 iterations and plotted the results against the sorted p-values from SGOC pairwise correlations separately for each cancer type (quantile-quantile plots).

Finally, mRNA expression data for the 8 colon cancer cell lines in study were obtained from Broad Institute Cancer Cell Line Encyclopedia (Barretina et al., 2012) and similar SGOC pathway expression analyses were performed on the data for comparisons with isotope labeling and flux data. Correlation between expression levels of SGOC genes with label ratios in Serine, Glycine, Thymidine, Glutathione, and IMP (representing purines) were calculated across the eight cell lines. As a control, label ratios were also correlated with expression of same-size sets of genes randomly picked from the genome. A histogram of all significant p-values ($p<0.05$) was generated from the results of the 500 simulations. Furthermore, gene expression at the pathway level (average across pathways shown) was also correlated with detected isotope enrichments. Single-genes and pathways were associated with fluxes by plotting the percent of significant correlations (p-value< 0.05) between gene expression and label ratios for each case. All calculations were carried out using R. (http://www.R-project.org/)

2.5.5 ^{13}C Mass-isotopomer distribution model

A two-compartment (mitochondria and cytosol) Serine-Glycine-Methylenetetrahydrofolate (Ser-Gly-meTHF) flux model was used to fit experimental ^{13}C mass isotopomer distributions (MIDs) of the SGOC associated metabolites to determine intercellular metabolic fluxes, transport and exchange fluxes relative to

extracellular serine transport flux. The model was formalized using two types of mass balance equations: 1) mass balance for total metabolite concentration; and 2) ^{13}C mass-isotopomer mass balance for labeled metabolites based on the network depicted in Figure 2.6A and corresponding atom distribution matrices. MID equations were derived in the similar manner as equations for bonded cumulative isotopomers as described previously (Shestov et al., 2012). In terms of the ordinary differential equations, the model describes the rates of loss and creation of particular labeled and unlabeled metabolite that forms after incubation of labeled serine in extracellular media. Those equations are based on the flux balance of metabolites and take the general form (e.g. for parallel unimolecular reactions):

$$[\boldsymbol{M}]\frac{\boldsymbol{d\mu_{(i)}}}{\boldsymbol{dt}} = \sum_{\boldsymbol{j}} \boldsymbol{F_j \sigma_{j(i)}} - \left(\sum_{\boldsymbol{k}} \boldsymbol{F_k}\right)\boldsymbol{\mu_{(i)}}$$

where metabolite M is downstream of another metabolites S_j. The total outflux $\sum F_k$ balances total influx $\sum F_j$. $[M]$ represents the total pool size of metabolite M, while $\mu_{(i)}$ and $\sigma_{j(i)}$ represent the I mass-isotopomer fraction of metabolite M ($M+I$ mass-isotopomer) and metabolite S_j ($S+I$ mass-isotopomer), respectively. The number of labeled C atoms in M molecule, I, changes between 0 and N, where N is the total number of C atoms in metabolites. At steady state the left term of equation (1) is equal to zero resulting in a set of algebraic equations. For labeled [U-^{13}C] serine experiments in RPMI medium, the fitted experimental steady state mass-isotopomers (combined cytosolic and mitochondrial) were four isotopomer forms of serine, three forms of glycine, four forms of glutathione which serves as a readout for cytosolic

glycine patterns, three forms of thymidine, and six isotopomer forms of adenine together making a total of 20 steady state mass-isotopomers. Thymidine and adenine isotopomers reflect pyrimidine and purine metabolism, respectively. Eleven fluxes were determined relative to serine transport flux, which are represented in Table A1.S1: de novo serine synthesis, two unidirectional rates for reversible cytosolic and mitochondrial SHMT fluxes, unidirectional glycine uptake flux, inter-compartmental serine, glycine, and one-carbon pool (meTHF) exchange fluxes, glycine cleavage system (GCS) activity, mitochondrial dilution flux for meTHF due to dimethylglycine and sarcosine contribution. Also the fraction of the de novo thymidine and adenine production along with salvage contribution was calculated. For each metabolite MIDs, ^{13}C natural abundance of ^{13}C isotope (1.08%) was taken into account. There was no need to correct for ^{15}N natural abundance (0.38%) due to the mass resolution used (70,000) that is able to separate ^{13}C and ^{15}N for the molecules considered. Solving a system of non-linear differential equations in terms of whole/fragmented mass-isotopomers, with the Runge-Kutta 4^{th} order procedure (Matlab, Natick, MA), yields time courses for all possible ^{13}C mass-isotopomers (e.g. serine and glycine). A quadratic cost function was used to quantify differences between measurements and estimated results for labeled steady state data and to select the corresponding vector of fluxes that minimizes the cost function in ^{13}C serine experiments using a simplex algorithm. Mean-square convergence was confirmed by verifying that goodness-of-fit values were close to expected theoretical values. To overcome potential local minima, we used several sets of initial random fluxes.

We estimated errors for the obtained values using a Monte Carlo simulation method as previously described (Shestov et al., 2007). Initial values for all parameters were chosen close to the HCT116 cell line fluxes as a reference cell line. ^{13}C mass-isotopomer values for Ser, Gly, GSH, Thd and Ade were then generated by solving differential equations describing the model with initial value of fluxes. For each Monte-Carlo run, random Gaussian noise with mean zero and standard deviation σ = 0.01 was added to the steady-state of these ^{13}C mass-isotopomers. Finally, the MIDs for each metabolite were computed and used for fitting with the SGOC metabolic model to obtain the values of relative metabolic fluxes. This procedure was repeated 500 times to obtain histograms for each parameter. Standard deviations are reported in Table A1.S1.

2.6. References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

Belmonte, M., Stasolla, C., Loukanina, N., Yeung, E.C., and Thorpe, T.A. (2003). Glutathione modulation of purine metabolism in cultured white spruce embryogenic tissue. Plant Sci *165*, 1377-1385.

Belmonte, M.F., Stasolla, C., Katahira, R., Loukanina, N., Yeung, E.C., and Thorpe, T.A. (2005). Glutathione-induced growth of embryogenic tissue of white spruce correlates with changes in pyrimidine nucleotide metabolism. Plant Sci *168*, 803-812.

Cancer Genome Atlas, N. (2012a). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330-337.

Cancer Genome Atlas, N. (2012b). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Cancer Genome Atlas Research, N. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

Chabner, B., and Longo, D.L. (2011). Cancer chemotherapy and biotherapy : principles and practice. (Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins).

Chaneton, B., Hillmann, P., Zheng, L., Martin, A.C., Maddocks, O.D., Chokkathukalam, A., Coyle, J.E., Jankevics, A., Holding, F.P., Vousden, K.H., et al. (2012). Serine is a natural ligand and allosteric activator of pyruvate kinase M2. Nature *491*, 458-462.

Circu, M.L., and Aw, T.Y. (2010). Reactive oxygen species, cellular redox systems, and apoptosis. Free radical biology & medicine *48*, 749-762.

Fan, J., Ye, J., Kamphorst, J.J., Shlomi, T., Thompson, C.B., and Rabinowitz, J.D. (2014). Quantitative flux analysis reveals folate-dependent NADPH production. Nature *510*, 298–302.

Gao, J., Tarcea, V.G., Karnovsky, A., Mirel, B.R., Weymouth, T.E., Beecher, C.W., Cavalcoli, J.D., Athey, B.D., Omenn, G.S., Burant, C.F., et al. (2010). Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics *26*, 971-973.

Gut, P., and Verdin, E. (2013). The nexus of chromatin regulation and intermediary metabolism. Nature *502*, 489-498.

Hu, C.M., Yeh, M.T., Tsao, N., Chen, C.W., Gao, Q.Z., Chang, C.Y., Lee, M.H., Fang, J.M., Sheu, S.Y., Lin, C.J., et al. (2012). Tumor cells require thymidylate kinase to prevent dUTP incorporation during DNA repair. Cancer cell *22*, 36-50.

Hu, J., Locasale, J.W., Bielas, J.H., O'Sullivan, J., Sheahan, K., Cantley, L.C., Vander Heiden, M.G., and Vitkup, D. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nature biotechnology *31*, 522-529.

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., Kafri, R., Kirschner, M.W., Clish, C.B., and Mootha, V.K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. Science *336*, 1040-1044.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research *28*, 27-30.

Labuschagne, C.F., van den Broek, N.J., Mackay, G.M., Vousden, K.H., and Maddocks, O.D. (2014). Serine, but Not Glycine, Supports One-Carbon Metabolism and Proliferation of Cancer Cells. Cell reports *7*, 1248–1258.
Lewis, C.A., Parker, S.J., Fiske, B.P., McCloskey, D., Gui, D.Y., Green, C.R., Vokes, N.I., Feist, A.M., Vander Heiden, M.G., and Metallo, C.M. (2014a). Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of Mammalian cells. Molecular cell *55*, 253-263.
Lewis, C.A., Parker, S.J., Fiske, B.P., McCloskey, D., Gui, D.Y., Green, C.R., Vokes, N.I., Feist, A.M., Vander Heiden, M.G., and Metallo, C.M. (2014b). Tracing Compartmentalized NADPH Metabolism in the Cytosol and Mitochondria of Mammalian Cells. Molecular cell.
Liu, X., Ser, Z., and Locasale, J.W. (2014). Development and quantitative evaluation of a high-resolution metabolomics technology. Analytical chemistry *86*, 2175-2184.
Locasale, J.W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. Nature reviews. Cancer *13*, 572-583.
Locasale, J.W., Grassian, A.R., Melman, T., Lyssiotis, C.A., Mattaini, K.R., Bass, A.J., Heffron, G., Metallo, C.M., Muranen, T., Sharfi, H., et al. (2011). Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. Nature genetics *43*, 869-874.
Ma, L., Tao, Y., Duran, A., Llado, V., Galvez, A., Barger, J.F., Castilla, E.A., Chen, J., Yajima, T., Porollo, A., et al. (2013). Control of Nutrient Stress-Induced Metabolic Reprogramming by PKCzeta in Tumorigenesis. Cell *152*, 599-611.
Maddocks, O.D., Berkers, C.R., Mason, S.M., Zheng, L., Blyth, K., Gottlieb, E., and Vousden, K.H. (2013). Serine starvation induces stress and p53-dependent metabolic remodelling in cancer cells. Nature *493*, 542-546.
Murphy, M.P., Holmgren, A., Larsson, N.G., Halliwell, B., Chang, C.J., Kalyanaraman, B., Rhee, S.G., Thornalley, P.J., Partridge, L., Gems, D., et al. (2011). Unraveling the biological roles of reactive oxygen species. Cell metabolism *13*, 361-366.
Nilsson, R., Jain, M., Madhusudhan, N., Sheppard, N.G., Strittmatter, L., Kampf, C., Huang, J., Asplund, A., and Mootha, V.K. (2014). Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. Nature communications *5*, 3128.
Possemato, R., Marks, K.M., Shaul, Y.D., Pacold, M.E., Kim, D., Birsoy, K., Sethumadhavan, S., Woo, H.K., Jang, H.G., Jha, A.K., et al. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature *476*, 346-350.
Scuoppo, C., Miething, C., Lindqvist, L., Reyes, J., Ruse, C., Appelmann, I., Yoon, S., Krasnitz, A., Teruya-Feldstein, J., Pappin, D., et al. (2012). A tumour suppressor network relying on the polyamine-hypusine axis. Nature *487*, 244-248.
Shestov, A.A., Valette, J., Deelchand, D.K., Ugurbil, K., and Henry, P.G. (2012). Metabolic modeling of dynamic brain (1)(3)C NMR multiplet data: concepts and simulations with a two-compartment neuronal-glial model. Neurochemical research *37*, 2388-2401.
Shestov, A.A., Valette, J., Ugurbil, K., and Henry, P.G. (2007). On the reliability of (13)C metabolic modeling with two-compartment neuronal-glial models. Journal of neuroscience research *85*, 3294-3303.
Shin, G., Kang, T.W., Yang, S., Baek, S.J., Jeong, Y.S., and Kim, S.Y. (2011). GENT: gene expression database of normal and tumor tissues. Cancer informatics *10*, 149-157.
Stasolla, C. (2010). Glutathione redox regulation of in vitro embryogenesis. Plant Physiol Bioch *48*, 319-327.
Tedeschi, P.M., Markert, E.K., Gounder, M., Lin, H., Dvorzhinski, D., Dolfi, S.C., Chan, L.L., Qiu, J., Dipaola, R.S., Hirshfield, K.M., et al. (2013). Contribution of serine, folate and

glycine metabolism to the ATP, NADPH and purine requirements of cancer cells. Cell death & disease *4*, e877.
Teperino, R., Schoonjans, K., and Auwerx, J. (2010). Histone methyl transferases and demethylases; can they link metabolism and transcription? Cell metabolism *12*, 321-327.
Vazquez, A., Markert, E.K., and Oltvai, Z.N. (2011). Serine biosynthesis with one carbon catabolism and the glycine cleavage system represents a novel pathway for ATP generation. PloS one *6*, e25881.
Vazquez, A., Tedeschi, P.M., and Bertino, J.R. (2013). Overexpression of the mitochondrial folate and glycine-serine pathway: a new determinant of methotrexate selectivity in tumors. Cancer research *73*, 478-482.
Wilson, P.M., LaBonte, M.J., Lenz, H.J., Mack, P.C., and Ladner, R.D. (2012). Inhibition of dUTPase induces synthetic lethality with thymidylate synthase-targeted therapies in non-small cell lung cancer. Molecular cancer therapeutics *11*, 616-628.
Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic acids research *30*, e15.
Zhang, W.C., Shyh-Chang, N., Yang, H., Rai, A., Umashankar, S., Ma, S., Soh, B.S., Sun, L.L., Tai, B.C., Nga, M.E., et al. (2012). Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. Cell *148*, 259-272.

CHAPTER 3: CONTRIBUTION OF ONE-CARBON METABOLISM TO DNA METHYLATION[3]

3.1 Abstract

Altered DNA methylation is common in cancer and often considered an early event in tumorigenesis. However, the sources of heterogeneity of DNA methylation among tumors remain poorly defined. Here, I capitalize on the availability of multi-platform data on thousands of human tumors to build integrative models of DNA methylation. I quantify the contribution of clinical and molecular factors in explaining within-cancer variability in DNA methylation. I show that a set of metabolic genes involved in the methionine cycle is predictive of several features of DNA methylation in tumors including the methylation of cancer genes. Finally, I demonstrate that patients whose DNA methylation can be predicted from the methionine cycle exhibited improved survival over cases where this regulation is disrupted. This study represents a comprehensive analysis of the determinants of methylation and demonstrates the surprisingly large association between metabolism and DNA methylation variation. Together, my results illustrate links between tumor metabolism and epigenetics and outline future clinical implications.

[3] Mehrmohamadi, M., Mentch K.L., Clark A.G., Locasale, J.W. Integrative modeling of tumor DNA methylation quantifies the contribution of metabolism. *Nature Communications*, (2016).

3.2 Introduction

DNA methylation is a major epigenetic mechanism that determines cellular outcome by regulating gene expression and chromatin organization(Jones, 2012) in a fashion more dynamic than previously appreciated (Schubeler, 2015). Altered DNA methylation is frequently observed in cancers compared to corresponding normal cells (Hansen et al., 2011; Mack et al., 2014; Timp and Feinberg, 2013). For example, global DNA hypomethylation (Ehrlich and Lacey, 2013) and tumor suppressor silencing by DNA hypermethylation are two of the most well characterized cancer associated alterations common across many human malignancies (Berman et al., 2012) . In addition to hypo- and hyper-methylation, cancer cells exhibit increased variability in DNA methylation across large portions of the genome compared to their corresponding normal tissues (Gaidatzis et al., 2014; Landau et al., 2014). Previous studies have shown that, for several cancer types, variation in methylation levels among tumor samples is significantly higher than normal samples of the same tissue of origin (Hansen et al., 2011; Landau et al., 2014), possibly indicating that deregulated epigenetics provides tumor cells with potential adaptive advantages (Timp and Feinberg, 2013). While inter-tissue variability in DNA methylation is mainly explained by differentiation and tissue-specific regulatory mechanisms (Lokk et al., 2014; Ziller et al., 2013), very little is known about the functions and determinants of the high inter-individual variation among tumors of the same tissue type. Notably, a recent twin study on the determinants of inter-individual variability in DNA methylation reported that genetic difference among individuals account for only 20%

of total variance with the remaining variance explained by environmental and stochastic factors that are yet to be identified(van Dongen et al., 2016).

The source of the methyl group for methylation is S-adenosylmethionine (SAM) which is generated from the methionine (met) cycle and is coupled to serine, glycine, one -carbon (SGOC) metabolism (Locasale, 2013). A large body of evidence indicates numerous roles for one-carbon metabolism in proliferation and survival of tumor cells through its roles in biosynthesis and redox metabolism (Gut and Verdin, 2013; Kaelin and McKnight, 2013; Locasale, 2013; Sahar and Sassone-Corsi, 2009). The met cycle also mediates histone and DNA methylation in physiological conditions and provides a link between intermediary metabolism and epigenetics (Anderson et al., 2012; Mentch et al., 2015; Pfalzer et al., 2014). Although the network contributes methyl units to DNA, whether and to what extent this interaction is apparent in tumors and may contribute to cancer biology is unknown.

I set out to comprehensively quantify the contribution of various factors in explaining variation in DNA methylation. The advent of standardized genomics and other high-dimensional multi-platform 'omics' data through The Cancer Genome Atlas (TCGA) allows for systematic assessments of molecular features across cancers (Cancer Genome Atlas Research et al., 2013). With combined statistical analysis, computational modeling, and machine-learning approaches, I directly evaluated the quantitative contributions of molecular and clinical variables that lead to DNA methylation. I found a surprisingly large contribution for the expression of the methionine cycle and related SGOC network genes in explaining DNA methylation

and identified numerous contexts where this interaction may contribute to cancer pathology.

3.3 Results

3.3.1 Quantification of the determinants of DNA methylation

It has been previously proposed that factors normally regulating the epigenome are disrupted in cancer, leading to increased variability of the cancer epigenome (Timp and Feinberg, 2013). However, the nature and contributions of such factors are largely unknown. Upon analysis of global and local DNA methylation in tumors as measured by the Illumina Infinium HumanMethylation450K BeadChip arrays, I indeed found higher variation among tumors from the same tissue vs. between different tissue types (Note A2.S1; Figure A2.S1A-D; Methods). Arrays were used over bisulfite sequencing because of the higher availability of these data in a standardized format allowing for an integrative analysis. To establish quantitative relationships between DNA methylation and molecular and clinical features of tumors, I developed an integrative statistical modeling and machine-learning approach with the goal of identifying the relative contributions to within-cancer DNA methylation variation (Methods). I incorporated hundreds of variables into comprehensive statistical models of DNA methylation (Figure 3.1A). Factors with a known role in DNA methylation machinery (chromatin remodeling enzymes and transcription factors), as well as factors with a potential biochemical link to DNA methylation (SAM-metabolizing

enzymes, met cycle enzymes, and other serine, glycine, one-carbon (other SGOC) enzymes that are connected to the met cycle (Mehrmohamadi et al., 2014)) were together considered (Figure 3.1A). I also curated available clinical information such as age, gender, and cancer stage in the calculations where appropriate. Furthermore, since mutations are known to affect the cancer methylome (Duncan et al., 2012), I included all recurrent genetic lesions (somatic mutations and copy number alterations) for each cancer type in my models. Together, over 200 variables were collectively analyzed for each cancer type. My models are therefore not completely agnostic as I pre-select classes of biological variables that are known to affect DNA methylation to avoid loss of statistical power by including too many features (e.g. expression of all genes in the genome). Therefore, to test for potential bias, I also considered the expression levels of sets of random genes with functions non-related to DNA methylation as additional variables in my models (see Methods). Subsequently, I incorporated all variables into unbiased selection algorithms suitable for dealing with large numbers of prediction variables. For this task, I considered two independent approaches: a generalized linear model (Elastic Net (Zou and Hastie, 2005)) and a machine-learning algorithm (Random Forest (Breiman, 2001)). A distinct computation was carried out for each 10 kilobase (kb) genomic region with variable methylation ($sd > 0.2$) in each cancer type. Samples of each cancer type were divided into three independent test subsets and three training subsets and separate models were generated using each subset. The models were then combined resulting in a single final model for each 10 kb region of DNA methylation in each cancer. Model

performance was evaluated by measuring mean squared prediction error of test samples from Elastic Net and Random Forest separately (Methods).

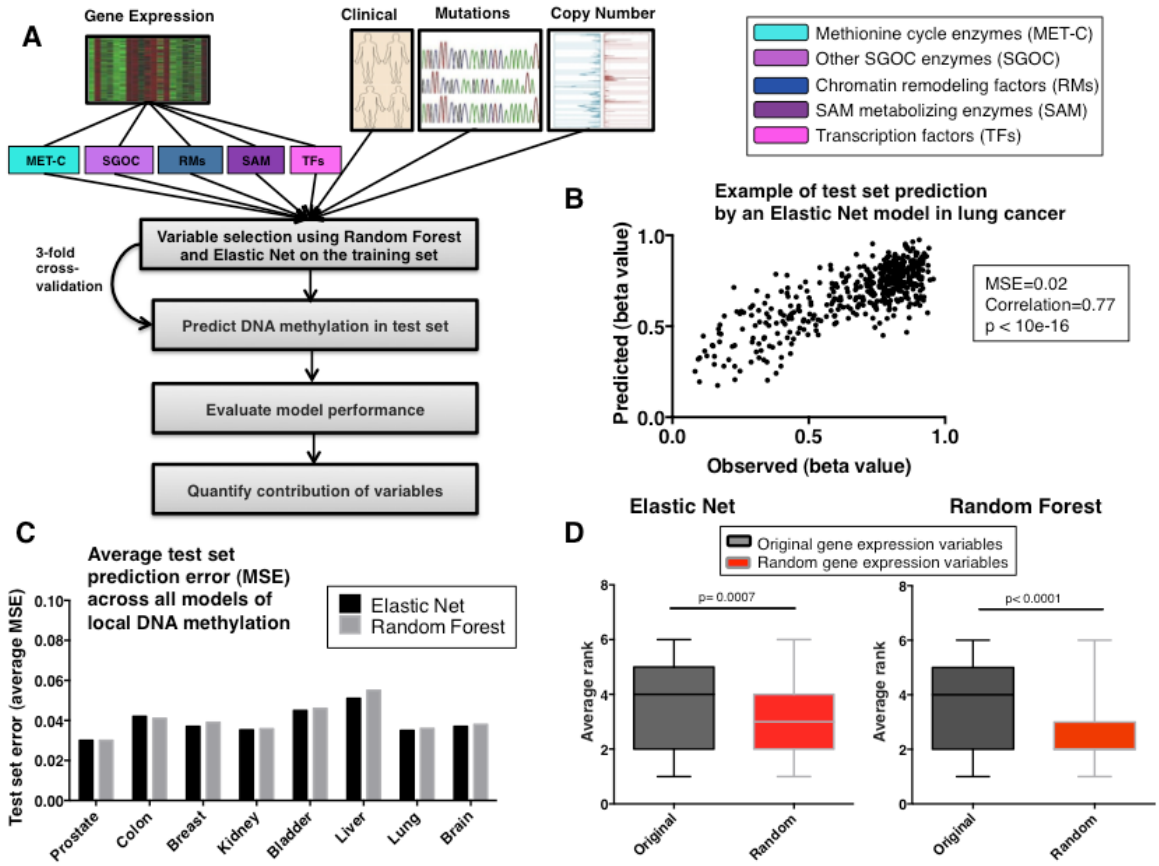


Figure 3.1—Integrative modeling of local DNA methylation levels.

A) Schematic summarizing the integrative approach utilized for modeling local DNA methylations. DNA methylation at a given 10 kb region was predicted by incorporating relevant gene expression, somatic mutation, copy number alteration, and clinical information into integrative models.

B) An example of an Elastic Net model performance in lung cancer. The x-axis shows true values of DNA methylation in each sample, and the y-axis shows the value predicted by the integrative modeling in the same sample when it was in the test subset.

C) Summary of overall model performance. For each cancer, the mean squared errors of test set predictions by Elastic Net and Random Forest were averaged across all models of local DNA methylation.

D) Comparison of original gene expression variables with randomly selected variance-matched genes. The y-axis shows the average rank of each gene expression category based on average variable usage score across all Elastic Net models (left) and average variable

importance score across all Random Forest models (right) of local DNA methylation in brain cancer (Error bars show the minimum and maximum value in each group). P-values associated with the Mann-Whitney test between the ranks across all models are shown (A higher rank corresponds to higher contribution; see Methods).

I observed that my models predicted test set DNA methylation with small mean squared error (MSE <0.04) in many regions across the genome (Figure 3.1B). Comparison of the performances of the two methods showed that Random Forest and Elastic Net algorithms were able to predict DNA methylation with comparable MSEs on average (Figure 3.1C; Figure A2.2A). In general, predictability of local DNA methylation was largely dependent on cancer type as well as chromatin region in each model. For example, I observed that local DNA methylation was most predictable in prostate and lung cancers and least predictable in liver and bladder cancers (Figure 3.1C; Figure A2.2A). Together with the high variation in local DNA methylation levels seen in liver and bladder cancers (Figure A2.1D), these results suggest a higher stochasticity in the epigenetic signatures for these two cancer types compared to others in this study. Upon annotating genomic regions where local DNA methylation could be predicted with a low error (MSE < 0.04) in each cancer, I found that the majority of the predictable regions lie within 20 kb of the transcriptional start site (TSS) of a gene (Figure A2.2B), suggesting that regulation of DNA methylation by the factors included in my models is stronger at genic regions.

I next performed a set of tests to evaluate the robustness of my modeling approach. To this end, I compared the original gene expression variables included in my models, with a group of variance-matched randomly selected genes from the

genome (see Methods) in their ability to predict DNA methylation. In the presence of both groups of gene expression variables (original and random), both Elastic Net and Random Forest models selected my original variables significantly more frequently than random genes (Higher rank corresponds to higher contribution; Mann-Whitney p-value= 0.0007 for Elastic Net and <0.0001 for Random Forest) (Figure 3.1D; see Methods). When the same test was performed in the presence of 5 additional popular gene families (Receptor tyrosine kinases (RTK), Receptor serine kinases (RSK), Toll like receptors (TLR), MAPK signaling (MAPK) and WNT signaling (WNT)), all but RTKs ranked significantly lower (Mann-Whitney p-value<0.0001) than the original gene expression variables that I initially included in my models based on biological functions (Figure A2.3; Method). Together, these tests validate my models and confirm that the Elastic Net and Random Forest algorithms are suitable for quantitation of variable contributions in determining DNA methylation. Given that my models are not completely agnostic, I do not rule out the possibility of existence of potentially highly contributing factors other than the hundreds of variables that I considered (e.g. RTKs). As such, the results should be interpreted in the context of relative contributions among the variables included in the models and the abilities of these variables in predicting DNA methylation.

3.3.2 Metabolism is a major predictor of DNA methylation in cancer

Using the results of the integrative modeling, I next quantified the relative contribution of different functional classes of variables in explaining DNA methylation variation within each cancer type. For this, I measured two independent

metrics, one using the Random Forest variable importance scores, and the other using a binary score for whether or not a variable was selected by the Elastic Net models (non-zero co-efficient). For each variable, an overall importance score was calculated by averaging its relative importance across all models of 10 kb DNA methylations, and an overall usage score was calculated by measuring the fraction of 10 kb regions in which Elastic Net models selected the variable (Methods). To estimate the contribution of each functional class of variables in explaining total variation in DNA methylation, I pooled all variables in the same functional category and averaged across their importance and usage scores separately (Figure A2.4A,B).

Results from both Random Forest and Elastic Net algorithms identified a considerable contribution from the variables within the SGOC metabolic network relative to other classes of variables ("Other SGOC enzymes" was the 2nd highest scoring among all classes, closely following "Transcription factors" according to both methods. "Methionine cycle enzymes" was the 3rd and 4th according to Random Forest and Elastic Net, respectively) (Figure 3.2A,B). Previous studies have shown that transcription factor abundance and occupancy strongly mediate dynamic DNA methylation turnover in regulatory regions (Feldmann et al., 2013; Stadler et al., 2011). Consistent with this observation, my results confirm the "Transcription factors" class has the highest contribution to predicting DNA methylation levels across human tumors. Notably, even in the presence of most if not all known variables that are thought to mediate the status of DNA methylation, metabolic factors still uniquely explained a large part of the variability in methylation (Figure 3.2A,B).

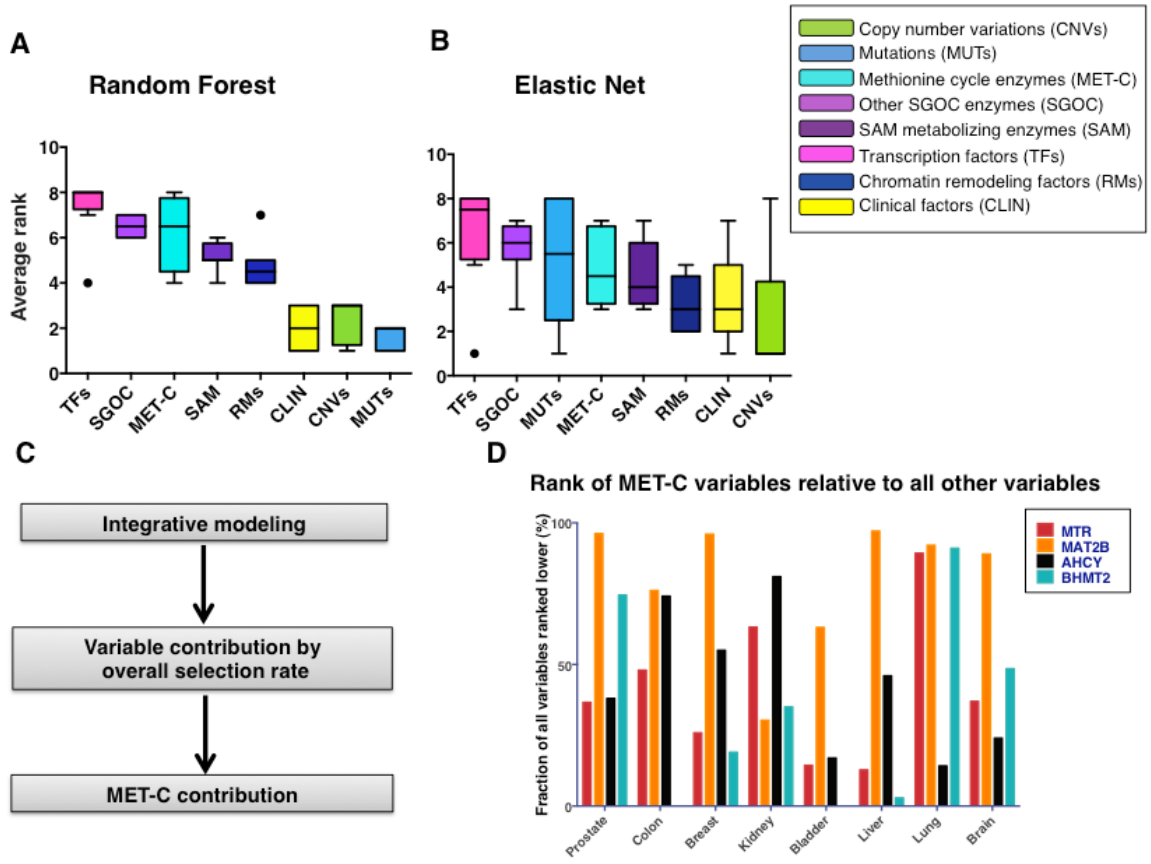


Figure 3.2—Contribution of different functional classes of variables to DNA methylation variation.

A-B) Relative contributions of the variable classes according to Random Forest average variable importance (A), and Elastic Net average variable usage (B) are shown averaged across all cancers (Methods). The y-axis shows the average rank of each class across cancers (with higher values corresponding to higher contribution). (Error bars span the range between the minimum and maximum values in each class with the exception of individual outliers shown).

C) Diagram summarizing the steps taken toward calculating overall contribution of each of the met cycle variables relative to other variables in explaining variability local DNA methylations.

D) Ranking all variables according to their overall selection rate (usage) across all models of local DNA methylation in each cancer. The y-axis shows the percent of variables that ranked lower than each of the met cycle variables (i.e. made less contribution to DNA methylation) in each cancer (BHMT2 was removed from the models of colon and bladder cancers due to low expression).

Given the contribution of the methionine cycle and its biochemical link to DNA methylation, I further explored the variables within the met cycle class compared to all other variables in their ability to predict DNA methylation (Figure 3.2C). Within the met cycle class, methionine adenosyltransferase 2 beta (MAT2B) and betaine-homocysteine S-methyltransferase 2 (BHMT2) exhibited higher predictive values than methionine synthase (MTR) and adenosylhomocysteinase (AHCY) on average (Figure A2.S4C,D). Notably, in the presence of the nearly 200 other variables in the computations, the met cycle — especially MAT2B— still contributed substantially to DNA methylation prediction (MAT2B was ranked among the top 5% of highly selected variables in prostate, breast, liver, lung, and brain cancers) (Figure 3.2D). I observed that the levels of MAT2B contribute to DNA methylation in nearly half of the variable regions across the genome even after accounting for various factors related to DNA methylation (MAT2B was selected by 42% of all Elastic Net models with MSE <0.04 on average) (Figure A2.S4C). Together, my results confirm that metabolism contributes to DNA methylation in many cases of human cancer and the association between metabolism and DNA methylation is stronger in some genomic regions than others.

3.3.3 Functional annotation of metabolically regulated regions

Results of the integrative modeling across cancers indicate that defined regulation of DNA methylation happens in regions where gene expression may be affected, thereby suggesting that this regulation could drive essential cancer biology. I next set out to characterize all regions across the genome where the association

between DNA methylation and the met cycle activity is particularly strong. To identify such regions, I designed a scanning algorithm to locate genomic regions spanning multiple CpGs with significant peaks of correlation of methylation with expression of met cycle enzymes (Figure 3.3A;Methods). I performed this analysis on each of the eight cancer types separately and identified distinct peak sets across the genome. To assess potential bias toward highly methylated regions and regions where there is higher probe density, I analyzed the relationship between average absolute methylation of individual CpGs and their correlation with met cycle expression, and found no significant association (p-value of correlation=0.62), confirming that the identified peaks are distinct from highly methylated regions (Figure 3.3B; Methods).

Density plots of peak distributions relative to the TSS of the nearest gene were concentrated around the TSS in all cancers (Figure A2.S5A), as expected given the higher density of probes in gene regulatory regions in the Illumina arrays (Figure A2.S5B). However, by further visualizing the distribution of the peaks immediately surrounding the TSS, I observed that peak distributions are more diffuse around the TSS (Figure A2.S5C) compared to the probe density distribution control (Figure A2.S5D). This suggests potential enrichment in areas of the genome overlapping with gene body regions and CpG island shores where dysregulated DNA methylation has previously been observed in human cancers (Timp and Feinberg, 2013). The peak distribution density plots extended up to a few hundred kilobases in distance from the nearest TSS, suggesting that DNA methylation at inter-genic parts of the genome may also be affected by the activity of met cycle.

I next tested the met cycle specificity of the identified peaks by correlating them with expression of randomly selected genes in the genome (Methods; Figure A2.S6A). For the majority (>83%) of the identified peaks, the met cycle's correlation with DNA methylation was significantly non-random (p-value<0.05) (Figure A2.S6B). These results show that my approach was able to identify genomic regions where DNA methylation levels are specifically affected by the met cycle activity.

I next set out to identify genes that overlap with the identified peaks in each cancer type. Functional annotation of genes overlapping these peaks by means of pathway enrichment analyses across a comprehensive collection of more than 70 gene-set libraries (Chen et al., 2013) showed enrichment of epigenetic features in these regions consistently across all cancers. Strikingly, many of my peaks overlapped with peaks of histone-3 lysine-27 tri-methylation (H3K27me3) (Figure 3.3C-F; Figure A2.S7A-D) as reported by both the encyclopedia of DNA elements (ENCODE) human project(Consortium, 2012) and the RoadMap epigenomics project (Roadmap Epigenomics et al., 2015). In cancers of the lung and bladder, histone-3 lysine-9 tri-methylation (H3K9me3) peaks were also significantly enriched (Figure 3.3F; Figure A2.S7C). H3K27me3 and H3K9me3 are both associated with repression of gene expression (Cedar and Bergman, 2009). My findings therefore suggest that variation in the met cycle's activity may contribute to aberrant expression from normally silenced loci and heterochromatin instability in cancer.

In addition to histone marks, tissue-specific and cell identity gene sets were also enriched in relevant cancer types, including "breast and ovarian cancer genes" in

breast cancer (Figure A2.S7A); “abnormal nervous system” and “abnormal neuron morphology” in brain cancer (Figure 3.3D); “asthma” and “lung carcinoma” gene sets in lung cancer (Figure 3.3F); “kidney-specific” gene set in kidney cancer (Figure A2.S7B); and “large intestinal genes”, “inflammatory bowel disease”, and “colorectal carcinoma” gene sets in colon cancer (Figure 3.3E). Finally, a number of developmental and signaling pathways were among the enriched pathways including “TGF-beta signaling” in kidney (Figure A2.S7B), “cell communication” pathway in liver (Figure 3.3C), and “G-protein coupled signaling” in bladder cancer (Figure A2.S7C). Organ and embryonic morphogenesis pathways were enriched in breast (Figure A2.S7A), bladder (Figure A2.S7C), and prostate (Figure A2.S7D), all of which are hormonally driven cancers. Interestingly, a previous study in breast cancer showed that embryonic developmental genes are enriched in regions of DNA hypomethylation compared to normal breast (Hon et al., 2012). Together, these results illustrate the functional importance of the relationship between met cycle and DNA methylation across cancers.

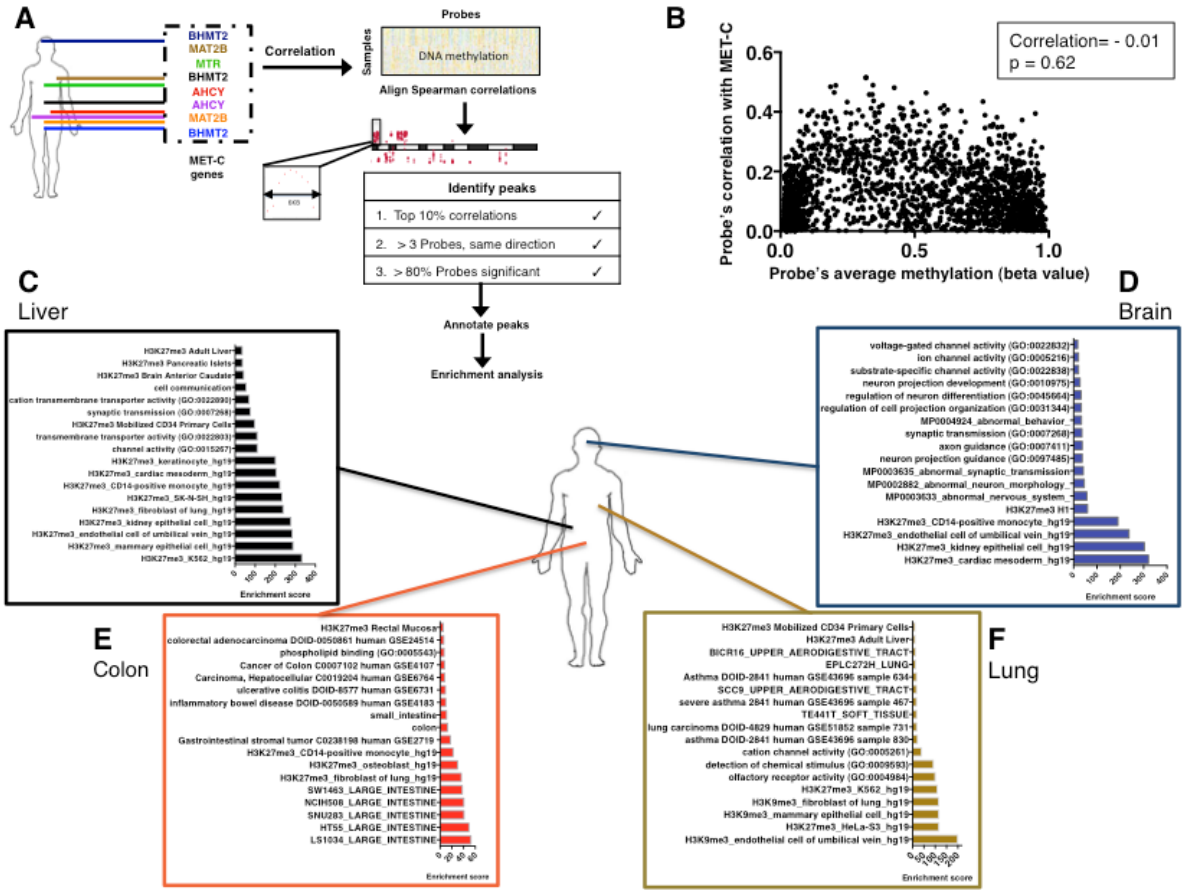


Figure 3.3— Genome-wide screening for metabolically regulated regions.

A) Schematic describing the algorithm used for finding genomic regions where DNA methylation might be regulated by the met cycle (see Methods).

B) Assessment of the relationship between met cycle correlation and absolute methylation. The y-axis shows the Spearman rho for correlation of 2000 randomly selected probes with the expression of AHCY in colon cancer. The x-axis shows the average methylation level of the same probes across the colon cancer samples in the study (see Methods).

C-F) Pathway enrichment analyses of genes overlapping peaks. Results are depicted by functional annotation of genes located within peaks of correlation between met cycle and DNA methylation (see Methods for description of gene sets and enrichment scores; see Figure A2.S7 for additional cancer types).

3.3.4 Contribution of metabolism to DNA methylation at cancer genes

So far, I have shown that there is a surprising enrichment of peak regions of metabolically regulated DNA methylation at loci that link to essential aspects of cell identity and chromatin structure. I next questioned whether cancer-specific loci might also exhibit this interaction. I chose 19 well-characterized cancer-related genes such as *TP53, PTEN,* and *ESR1*, as well as 4 genes frequently differentially methylated in cancer *APC, RASSF1, GSTP1,* and *MGMT* (Methods). A recent study showed that DNA methylation for any given gene has two major principal components: one representing the promoter region and the other representing the coding sequence (Ho et al., 2015). Furthermore, CpG methylation at promoter regions of genes is typically associated with repression, while gene body methylation is thought to increase expression (Yang et al., 2014). I therefore applied my integrative modeling to DNA methylation at promoter and gene body regions of each cancer gene separately. In addition to the integrative approach, I also generated models using only the met cycle genes as prediction variables to quantify the predictive ability of met cycle in the absence of other factors. Thus, each cancer gene locus was analyzed once using the integrative approach and once using met cycle alone and 3-fold cross validation was performed in each case as previously described (Methods). Model performance was evaluated by calculating the error of prediction of test set methylation, as shown for two examples in Figure 3.4: estrogen receptor (*ESR1*) promoter in breast cancer (MSE=0.004) (Figure 3.4A), and androgen receptor (*AR*) promoter in prostate cancer (MSE= 0.001) (Figure 3.4B). *ESR1* promoter methylation in breast cancer and *AR* promoter methylation in prostate cancer are two examples of events that are known to contribute to the pathogenesis and prognosis of the corresponding tumor types (Heyn

and Esteller, 2012; Nakayama et al., 2000; Yang and Park, 2012). I further assessed the integrative models of promoter methylation at these two loci, and found many SGOC (including met cycle) variables among the top predictive variables of promoter methylation according to the variable importance measures (Figure 3.4C,D; Methods).

Notably, the models across all cancers in the study were able to predict cancer gene methylation with high accuracy even using the met cycle variables in the absence of all other variables (85% of the predictions were made with MSE<0.01) (Figure A2.S8A,B). As in the case of local methylation, cancer gene methylation was also more strongly explained by the expression of MAT2B compared with other met cycle variables on average (selected by 24% of all integrative models) (Figure A2.S8C), consistent with the function of this enzyme that directly affects SAM levels. Relative variable class comparisons confirmed considerable contribution from the "Methionine cycle enzymes" and "Other SGOC enzymes" among other classes of variables (highest after "Transcription factors" and "Mutations") (Figure A2.S8D,E).

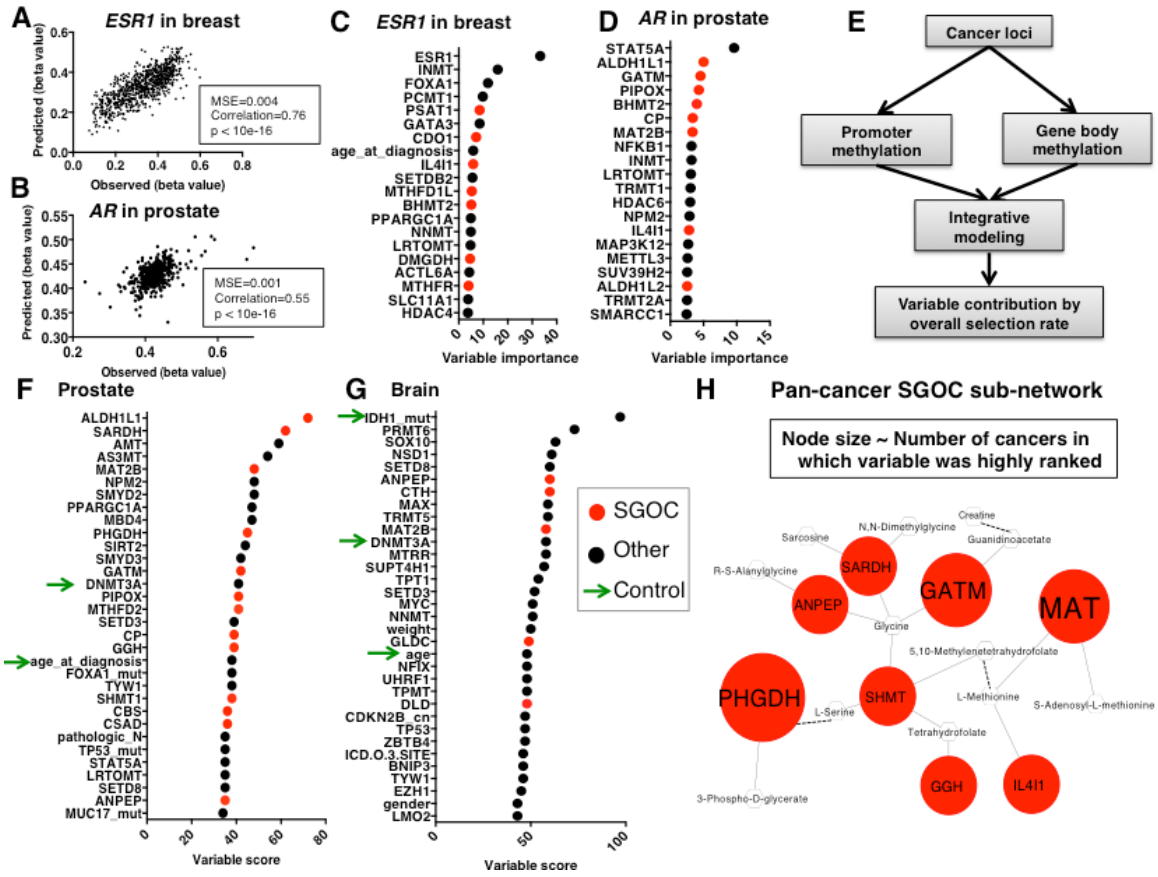


Figure 3.4— Contribution of metabolism to DNA methylation at cancer loci.

A) Prediction of ESR1 promoter methylation in test samples of breast cancer. The x-axis shows the methylation value at ESR1 promoter, while the y-axis shows the corresponding predicted values by Elastic Net.

B) Prediction of AR promoter methylation in test samples in prostate cancer. The axes are similar to a.

C-D) Top 20 variables as ranked based on the variable importance score from Random Forest model of ESR1 promoter methylation in breast cancer (C), and AR promoter methylation in prostate cancer (D). Variables in the SGOC network (including the met cycle enzymes and other SGOC enzymes) are shown in red and all other variables are shown in black.

E) Schematic depicting the ranking of all variables based on combined results of promoter and gene body methylation at cancer loci.

F-G) Variables that were most predictive of cancer gene methylation on average (top 15%) are ranked in order of increasing contribution (variable score= percent usage by Elastic Net). Green arrows point to previously published factors associated with variations in DNA methylation (positive controls). (Variable names: official gene symbols are used to show gene expression variables ("Methionine cycle enzymes", "Other SGOC enzymes", "Transcription factors", "Chromatin remodeling factors", and " SAM metabolizing enzymes"), while

"_mut" and "_cn" suffixes following gene symbols denote "Mutations" and "Copy number variations", respectively. For "Clinical factors", variable names match the descriptors used in the TCGA data files) (see Figure A2.S11 for additional cancer types).

H) Sub-network of SGOC genes contributing to DNA methylation in multiple cancer types (at least 4 and 3 cancers based on Elastic Net and Random Forests models, respectively). Red and white nodes represent genes and metabolite, respectively. Solid edges denote direct biochemical links and dashed edges denote indirect biochemical links through enzymatic reactions not shown. Node sizes for the gene nodes correspond to the number of cancer types wherein each enzyme contributed significantly to cancer gene methylation. (Phosphoglycerate dehydrogenase (PHGDH)=6, MAT (MAT2B and MAT2A) =5, glycine amidinotransferase (GATM)=5, serine hydroxymethyltransferase 1 and 2 (SHMT1 and SHMT2)= 4, sarcosine dehydrogenase (SARDH)= 4, alanyl aminopeptidase (ANPEP)= 4, L-amino acid oxidase (IL4I1)=4, gamma-glutamyl hydrolase (GGH)=4).

I independently evaluated these findings by applying the same models to both permuted cancer gene methylation values and also randomly generated methylation values (Figure A2.S9A). In all tests, met cycle contribution was significantly (p-value< 10e-16) higher when applied to cancer gene methylation vs. permutations or random numbers (Methods; Figure A2.S9B), confirming the specificity of signals contained in the true DNA methylation values at cancer loci. Furthermore, my collaborator Lucas Mentch tested the performance of the machine-learning algorithm using randomly generated variables for prediction of cancer gene methylation (Methods) and found in each of the cases tested, that the predictions made with the original variables are uniformly more accurate than what is made using simulated random variables (original model MSE smaller by 1.4-2 fold than random model MSE on average) (Figure A2.S10A-D). He also simulated a dataset where prediction variables and the response are related via linear relationships and compared the accuracy of predictions in this simulated linear dataset with the original dataset (Methods). He saw in all cases that the improvement in MSE from my dataset (MSE

1.4-2 fold smaller than random MSE) is even more than what we observed with data of the same dimension that have a linear relationship (MSE 1.3 fold smaller than random MSE) (Figure A2.S10E,F). These independent tests confirm that machine-learning using the Random Forest algorithm is able to identity non-random signals in the data, and also that it can detect non-linear relationships between prediction variables and the response.

Next, I ranked all of the variables based on their overall usage according to the integrative models of cancer gene promoter and body methylations (Figure 3.4E). Notably, many SGOC (including met cycle) enzymes were among the most frequently selected variables in all cancers (Figure 3.4F,G; Figure A2.S11A-F). Importantly, my models highly ranked many clinical and molecular factors previously shown to be associated with DNA methylation in the existing literature (green arrows in Figure 3.4F,G and Figure A2.S11A-F). Examples of such positive controls include DNA methyltransferase (DNMT3A or DNMT3B) enzymes (Robertson, 2001) that were consistently among the top variables in all cancers (Figure 3.4F,G; Figure A2.S11A-F), and patient's age (or age at diagnosis) (Jung and Pfeifer, 2015; van Dongen et al., 2016) that was highly ranked in prostate, colon, breast, kidney, and brain (Figure 3.4F,G; Figure A2.S11A-F). I also observed ER-status to be one of the most important contributors to DNA methylation variation in breast cancer consistent with previous publications (Cancer Genome Atlas, 2012) (Figure A2.S11B). Furthermore, I found the mutational status of the histone methyltransferase SET-domain containing-2 (*SETD-2*) as a significant contributor in kidney (Figure A2.S11C), smoking in bladder and lung (Figure A2.S11D,F), and isocitrate dehydrogenase-1 (*IDH1*) mutational

status in brain cancers (Figure 3.4G). Each of these findings are in agreement with the current knowledge about determinants of DNA methylation (Cancer Genome Atlas Research, 2013, 2014; Turcan et al., 2012; van Dongen et al., 2016). These results further validate my models and also emphasize the importance of the contribution observed for the SGOC variables (including the met cycle).

Previous work has shown that expression of enzymes across different regions of the SGOC network is predictive of metabolic flux through the network (Mehrmohamadi et al., 2014). Notably, I observed that several SGOC genes are consistently among the highly ranked variables by both Random Forest and Elastic Net models in multiple cancer types. Therefore, to understand which features of SGOC metabolism contribute to the interaction with methylation, I defined a sub-network that was commonly highly ranked by the models in multiple cancer types (Figure 3.4H; Methods). This SGOC sub-network comprises the MAT enzymes in the met cycle (MAT2B and MAT2A), as well as enzymes within serine-glycine metabolism such as phosphoglycerate dehydrogenase (PHGDH) and glycine amidinotransferase (GATM) (Figure 3.4H). I generally observed negative associations between DNA methylation and expression of PHGDH and GATM, but positive associations with expression of MAT enzymes. A cautionary note however is that in many disease states, levels of particular metabolites in the methionine cycle substantially deviate from physiological ranges, thus activating compensatory mechanism and leading to correlation with DNA methylation in directions opposite of what would be expected from the biochemistry of the reactions (Jia et al., 2016). Therefore, when interpreting the direction of correlations between metabolic enzyme

levels and DNA methylation, it is important to note that they not only depend on the stoichiometry of the corresponding enzymatic reactions, but also on endogenous abundance of the related metabolites. Together, my results suggest that a particular flux configuration through the SGOC metabolic network— which previous studies have shown to be predictable from gene expression patterns (Mehrmohamadi et al., 2014)—may be important for regulation of DNA methylation.

3.3.5 Cancer pathogenesis of metabolically regulated DNA methylation

Involvement of the met cycle in promoter and gene body methylation at cancer genes suggests a potential implication for this metabolic pathway in explaining part of the variability in cancer pathogenesis and patient outcome. To further assess this relationship, I divided patients in each cancer type into two groups based on overall predictability of their cancer loci methylation by the met cycle (see Methods). I then compared survival rates between the two groups ("predictable" by met cycle vs. "not predictable" by met cycle) in each cancer type using the Kaplan-Meier estimator (Kaplan and Meier, 1958) (Figure 3.5A-H). An improved overall survival for the "predictable" group was observed, although the magnitude of this trend varied depending on cancer type with brain, kidney, liver, and colon cancers showing statistically significant differences (log-rank test p-values: brain= 3.92e-05, liver=0.0048, kidney=0.0085, colon=0.04) (Figure 3.5A-D). The difference in survival between the predictable and non-predictable groups was not significant in the rest of the cancers studied here (Figure 3.5E-H), possibly explained by limited power due to data censoring at later time points. The overall patterns however suggest that the

regulation of DNA methylation by the met cycle may be important in maintaining a normal epigenome, and disruption of this relationship in specific subtypes of tumors can lead to high epigenetic stochasticity in those tumors that correspond to poor clinical outcomes. This is consistent with a previous study that showed DNA methylation stochasticity increased across samples with increasing malignancy (from normal to adenoma to carcinoma) (Timp and Feinberg, 2013).

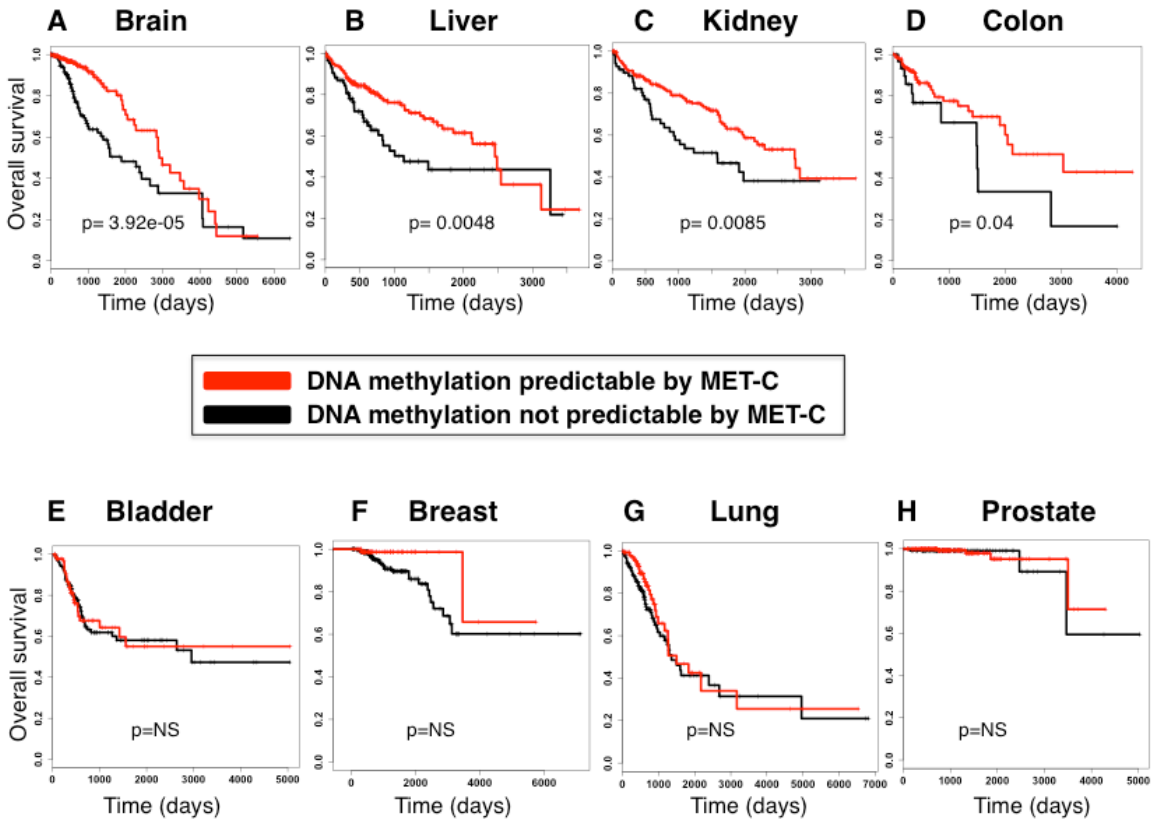


Figure 3.5— Implication of metabolic regulation of methylation in patient survival.

A-D) Kaplan-Meier curves are depicted comparing groups of patients wherein cancer gene methylation was predictable (red) or not predictable (black) by the met cycle variables (see Methods). Overall survival in days is plotted in each case and censored subjects are shown by vertical tick marks (Methods). Log-rank test p-value between the two groups is reported. Survival analysis results and log-rank test p-values are shown for brain, liver, kidney, and colon cancers respectively.

E-H) Survival analysis results as described above are reported for bladder, breast, lung, and prostate cancers, respectively. Log-rank test p-values showed no significant difference between the "predictable" and "not predictable" groups ("NS"= not significant) (Sample sizes: breast=770, lung=450, liver= 374, brain= 534, bladder= 408, kidney=316, prostate= 424, colon= 198).

To validate the results of my survival analyses, I applied multivariate Cox regression models to account for covariates such as mutations and clinical factors that are known to be associated with survival rates (Methods). I performed this test in the cases of brain, liver, and kidney cancers were the univariate analyses found highly significant differences between the predictable and non-predictable groups (Figure 3.5A-C). The models including covariates still showed a significant difference ($p<0.05$) between the predictable and non-predictable groups of patients even after taking mutational and clinical factors into account (see Methods for the list of covariates considered in each cancer), suggesting that a unique part of variation in survival may be explained by epigenetic regulation. I next tried to further validate my results through comparison with independent analyses of the TCGA data by the cBioPortal for Cancer Genomics (cBioPortal) (Cerami et al., 2012) and Prediction of Clinical Outcome from Genomic profiles (PRECOG) (Gentles et al., 2015). These analyses found lower survival in prostate cancer patients harboring tumors with deep deletions in the met cycle genes (Figure A2.S12A), and higher survival in kidney cancer patients where the met cycle enzymes are over-expressed (Figure A2.S12B), respectively. These results confirm a relationship between met cycle and survival in the same direction as predicted by my hypothesis.

3.4. Discussion

In this study, I conducted a pan-cancer TCGA analysis of the molecular and clinical contributions to within-cancer (inter-individual) variation in DNA methylation. Through several lines of integrated analysis, I found the overall expression of both the methionine cycle and SGOC network to be strong predictors of multiple aspects of DNA methylation and consistently ranked as one of the highest contributing factors to cancer-associated DNA methylation such as methylation of numerous cancer genes. Within the methionine cycle, I consistently observed a more significant contribution from MAT2B and BHMT2, suggesting that the regulation may be occurring at these enzymatic steps. MAT2B is the enzyme that converts methionine to SAM, therefore it is expected that this enzyme affects SAM levels more directly than other metabolic enzymes. The significance of BHMT2 but not MTR suggests that metabolism of choline and betaine may be more prevalent than folates in cases where one-carbon metabolism fuels DNA methylation. It is important to note that given the predictive nature of my models, the results do not prove causal relationships. As such, they should not be interpreted as direct evidence for regulation of DNA methylation by the model variables, but rather as predictive associations.

I introduced a novel approach to identify chromatin regions with strong correlations between DNA methylation and metabolic enzyme levels. The identified regions for the met cycle enzymes significantly overlapped with histone modifications, consistent with enzymatic cross-talk between the two epigenetic processes (Cedar and Bergman, 2009). The enrichment of gene signatures of repressing histone marks such

as H3K27me3 in all cancers points to a possible role for the met cycle in maintenance of DNA methylation at silenced loci. Previous studies have reported aberrant methylation of transcriptionally repressed genes in cancer (Sproul et al., 2011). In fact, heterochromatin instability arising from increased variability in DNA methylation is a phenomenon observed in many cancers and is thought to contribute to epigenetic plasticity and tumor progression (Carone and Lawrence, 2013; Hansen et al., 2011; Landau et al., 2014). My results provide evidence for this model of dysregulated cancer epigenome and further suggest that disruption of the regulation of DNA methylation by the met cycle—which can be a cause or consequence of tumorigenesis— may be one of the sources of methylation stochasticity leading to higher malignancy. Survival analyses confirm that tumors with a weaker association between their cancer gene methylation and the met cycle expression are more malignant in comparison to tumors wherein this relationship is closer to normal. In addition to epigenetic overlaps, genes with important tissue-specific functions and disease states were also found to fall under the metabolism-DNA methylation peaks. DNA methylation at cell-type related disease and lineage-specific genes has previously been shown to be dynamic and functionally important (Ziller et al., 2013). My results further strengthen the idea that met cycle regulation of methylation is strongly associated with normal tissue function.

Application of the integrative modeling to cancer genes revealed a major role for MAT enzymes (MAT2B and MAT2A), as well as PHGDH and GATM— enzymes involved in serine and glycine metabolism, respectively. Importantly, MAT2B and MAT2A have been shown to co-localize in nuclei and bind DNA through complex

formation with chromatin binding proteins providing direct evidence for the role of these enzymes in regulation of transcription via methylation (Katoh et al., 2011). My results illustrate that higher levels MAT2B are associated with more "regulated" methylation and higher survival, suggesting potentials for genetic or dietary interventions with methionine cycle intermediates in cancer patients. PHGDH diverts the glycolytic flux into the de novo serine synthesis pathway that allows glycolysis to provide methyl units. GATM diverts glycine into the creatine synthesis pathway in which SAM is consumed to produce creatine (Brosnan et al., 2011). Creatine synthesis is therefore in competition with the methionine cycle over cellular pools of SAM, explaining why enzymes within the serine-glycine metabolism generally tend to be negatively correlated with the met cycle and DNA methylation.

Overall, this study provides the first comprehensive quantification of the determinants of inter-individual DNA methylation variation in human cancers. The activity of the methionine cycle that emerges in these findings could be either sensed directly by the DNA, or indirectly through interplay with dynamic histone methylation, which itself is tightly regulated by the status of methionine metabolism (Mentch et al., 2015). Due to limitation in the coverage of the DNA methylation arrays, it remains to be determined if my findings are generalizable to methylation across the entire genome including all non-CpG methylation sites as well as hydroxy-methylation sites. Nevertheless these findings altogether identify metabolism as a major determinant of DNA methylation status in human cancer. It is important to note that the current TCGA dataset contains one sample per individual tumor and therefore my conclusions do not necessarily explain the variation in clonal populations within a

given tumor. Future studies using multiple samples per tumor or single cell epigenomics are therefore required to characterize the determinants of intra-tumor epigenetic heterogeneity. Finally, my study identifies an association between altered tumor metabolism and DNA methylation, while the sources of alterations in metabolism itself remain to be elucidated but can be addressed using similar approaches.

3.5 Methods

3.5.1 Data curation

Publically available genome-wide mRNA expression and DNA methylation data were downloaded from the cancer genome atlas (TCGA) portal *(https://tcga-data.nci.nih.gov/tcga/)*. In order to increase consistency and minimize unwanted variations, only samples processed using RNASEQ-V2 with level-3 gene-normalized RNA-seq by Expectation Maximization (RSEM) values for gene expression, and level-3 beta-values from Illumina Infinium HumanMethylation450K BeadChip data for DNA methylation were included in the study. I selected the following 8 cancer types wherein the number of available samples analyzed on both platforms was sufficiently large for machine-learning calculations: 770 samples of breast invasive carcinoma (BRCA), 450 samples of lung adenocarcinoma (LUAD), 374 samples of liver hepatocellular carcinoma (LIHC), 534 samples of brain lower grade glioma (LGG), 408 samples of bladder urothelial carcinoma (BLCA), 316 samples of kidney

renal clear cell carcinoma (KIRC), 424 samples of prostate adenocarcinoma (PRAD), and 198 samples of colon adenocarcinoma (COAD). Somatic mutations with a frequency of 5% or higher, and Genomic Identification of Significant Targets in Cancer (GISTIC) values for copy number alterations with a frequency of 15% or higher according to the cBioPortal (Cerami et al., 2012) were obtained and included in the models. Clinical and follow-up data were downloaded via the TCGA-Assembler (Zhu et al., 2014) .

3.5.2 Assessment of batch effects

I used the TCGA Batch Effects online tool (http://bioinformatics.mdanderson.org/tcgabatcheffects) to check for the existence of batch effects in the data used in my study. For each cancer types in my study, both the DNA methylation and the RNA-seq batch effects were negligible (Dispersion Separability Criterion (DSC) score < 0.5 for all sample batches included in the study).

3.5.3 DNA methylation

The Illumina Infinium Human Methylation450K BeadChip consists of more than 450,000 probes across the genome covering CpG sites within and outside of CpG islands as well as non-CpG methylation sites identified in embryonic stem cells *(see: http://www.illumina.com/products/methylation_450_beadchip_kits.html)*. I first filtered all probes with more than 80% missing values across each cancer type. Global

DNA methylation was then defined as the average beta-value across all remaining probes for each sample (Figure A2.S1A). Sex chromosomes were also excluded from all subsequent analyses of DNA methylation. In order to assess local DNA methylation, I divided the genome into 10 kb intervals and calculated the average beta value across all probes within each bin. I then filtered regions where variation in methylation was modest (standard deviation < 0.2 across each dataset). The average beta-value across all remaining 10 kb regions was then calculated for each sample individually and plotted in Figure A2.S1C. In order to study DNA methylation at cancer loci, probes that mapped to each gene according to Illumina annotations were identified. Promoter DNA methylation was then defined as the average beta value across all probes mapping to a given gene and falling within one of the following positional categories based on Illumina chip annotation information: "TSS1500", "TSS200", or "5'UTR". Gene body methylation for each gene was defined as the average beta value across all probes mapping to a given gene and falling in "1st exon", "Body", or "3'UTR" based on the annotation. Promoter and gene body methylation were separately modeled for each of the cancer genes in the study (Figure 3.4).

3.5.4 Gene expression

Log-transformed gene normalized RSEM values were used as expression levels and low-expression genes in each dataset were defined as having less than 70% of the samples with a count value larger than 3. Such genes were removed from further analysis.

3.5.5 Gene expression variables included in the integrative models

In addition to the major enzymes in the met cycle (MAT2B, MTR, BHMT2, AHCY), four classes of expression variables with potential links to DNA methylation were also included in the integrative models . The four classes are described in the following:

"Other SGOC enzymes": Serine, glycine, one-carbon (SGOC) metabolic genes from my previous network reconstruction were included (Mehrmohamadi et al., 2014). In order to separately assess the effect of the met cycle from the rest of the network, I excluded the met cycle enzymes from this class and treated them as a separate class ("Methionine cycle enzymes").

"Chromatin Remodelers": A list of human chromatin remodelers and DNA methylation machinery was constructed by combining the Gene Ontology (GO) chromatin modifiers list, GO chromatin remodelers list (Ashburner et al., 2000), and methylated DNA binding proteins and de-methylases (Marchal and Miotto, 2015).

"Transcription factors": For each cancer type, transcription factors important in the pathogenesis or subtype specification based on previous literature were included (Johnston and Carroll, 2015).

"SAM-metabolizing enzymes": DNA methyltransferases and other SAM-consuming enzymes (except for MAT enzymes already included in the class "Methionine cycle enzymes") according to Human Cyc (Romero et al., 2005) were included in this class.

3.5.6 Mutations included in the integrative models

For each cancer type, genes with frequent somatic mutations (minimum frequency of 5%) among the TCGA cohort according to the cBioPortal (Cerami et al., 2012) summary table (TCGA, Provisional) were obtained. The transposed matrix of individual barcodes and mutations in the selected genes was downloaded from the cBioPortal for each of the 8 cancers in this study.

3.5.7 Copy number alterations included in the integrative models

For each cancer type, genes with frequent copy number alterations (minimum frequency of 15%) among the TCGA cohort according to the cBioPortal (Cerami et al., 2012) summary table (TCGA, Provisional) were obtained. The transposed matrix of individual barcodes and putative copy number alteration calls by GISTIC (Mermel et al., 2011) for the selected genes was downloaded from the cBioPortal for each of the 8 cancers in this study (Values of putative copy number calls determined using GISTIC 2.0 : -2 = homozygous deletion; -1 = hemizygous deletion; 0 = neutral / no change; 1 = gain; 2 = high-level amplification).

3.5.8 Clinical factors included in the integrative models

For each cancer type, clinical information was downloaded through the TCGA-Assembler (Zhu et al., 2014). All clinical attributes were included for each cancer type

with the exception of the ones filtered out due to missing data for all samples or factors with the same level across all samples.

3.5.9 Variable ranking using the Random Forest algorithm

The Random Forest is a machine-learning algorithm that generates predictions by averaging over a collection of randomized decision trees. Since successive trees are built with bootstrap samples, the algorithm is robust to over-fitting, and also those samples that are left out (the *out-of-bag* (OOB) samples) can be used to quantify the contribution that prediction variables make to the overall response. The Random Forest method is designed to accommodate nonlinearities between the response and prediction variables as well as unknown interactions among the variables (Costello et al., 2014; Mentch and Hooker, 2015). I used the R package "randomForest"(Liaw and Wiener, 2002) and performed 3-fold cross validation by manually dividing the samples in each cancer type into 3 training and test subsets. To build each forest, tree size was set to 500 and the "importance" parameter was set to "TRUE" in the R function "randomForest" so as to provide estimates for the importance of prediction variables. Missing data were imputed using the "na.roughfix" function in the "randomForest" package. I obtained separate measures of importance for each variable from each Random Forest run. These importance scores are calculated as the percent increase in the mean squared prediction error on the OOB samples when a given variable is permutated. Variables were ranked based on average importance scores across all cross validation folds. Prediction errors were calculated as the mean squared difference between the predicted vs. the observed methylation values for the

test set samples. The square root of the mean squared error (MSE) has the same scale as the response (DNA methylation beta values in this case), and is therefore a direct measure of the accuracy with which predictions were made. (Figure 3.1C; Figure A2.S8A,B).

3.5.10 Variable selection using the Elastic Net algorithm

Elastic Net is a penalized regression approach for variable selection and quantitative inference that identifies linear combinations of unique variables that contribute to a response variable such as the amount of DNA methylation. The algorithm was developed and benchmarked to avoid over-fitting in statistical modeling of high-dimensional data containing collinearity (Waldmann et al., 2013). I applied the Elastic Net algorithm using the R package "glmnet"(Friedman et al., 2010). Elastic Net performs variable selection by minimizing a regularized cost function using the following equation

$$\boldsymbol{min}_{\boldsymbol{\beta 0,\beta}} \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{w_i l}(\boldsymbol{y_i}\, \boldsymbol{\beta^T x_i}) +) + \lambda\, [(1-\alpha)\, ||\boldsymbol{\beta}||_{\mathbf{2}}^{\mathbf{2}}/2 + \boldsymbol{\alpha}||\boldsymbol{\beta}||_{\mathbf{1}}]$$

where lambda is the tuning parameter and alpha is the Elastic Net penalty term. For each cancer type, the samples were divided into 3 independent test subsets (3-fold cross validation), and separate models were generated using each training subset. Using a grid of different tuning parameter values, I found the lambda that minimized the mean squared error using 5-fold cross validation within each training set for each model separately. The value of alpha was set to α=0.5 to handle potential correlated variables. Finally, for each variable, average coefficient across the 3 independent

models was calculated for each region and each cancer type. Due to the existence of categorical factors among my variables (for which scaling is not appropriate), I also calculated the selection rate as an alternative measure of variable importance referred to as "variable usage" in the manuscript. Variable usage was measured as the fraction of times across all cross validation folds that a variable was selected by the Elastic Net to be included in the final model (Figure A2.S8C; Figure 3.4 F,G; Figure A2.S11A-F). Finally, prediction errors were calculated as the squared difference (mean squared error (MSE)) between the predicted and measured DNA methylation values for the test sets (Figure 3.1C; Figure A2.S8A,B).

3.5.11 Variable class contributions to DNA methylation

Variables were functionally categorized into the following 8 classes: "Methionine cycle enzymes", "Other SGOC enzymes", "Chromatin remodeling factors", "Transcription factors", "SAM metabolizing enzymes", "Clinical factors", "Copy number variations", and "Mutations". Results of the integrative modeling were summarized and reported in terms of the average contribution from each of the above functional classes in explaining DNA methylation variation. Variable importance scores from Random Forest models were averaged across all variables within a given class, and an overall class importance score was calculated. In the case of Elastic Net models, variable usage as described in the previous section, was averaged across variables in each class and an average percentage showing selection rate was calculated. Finally, classes were ranked in each cancer type according to their average

contribution and the overall class ranks were plotted in Figure 3.2A,B, Figure A2.S4A,B, Figure A2.S8E,D.

3.5.12 Comparison with gene expression controls

A set of 100 randomly selected genes from the genome with similar cross-sample variation in expression as my original gene expression variables (TFs, SGOC, MET-C, SAM, and RMs) were considered. I performed this test on local DNA methylations (all variable 10 kb regions) in brain cancer (LGG) as an example and repeated the integrative modeling using this set of randomly selected genes in addition to all other variables present in the original models. All gene expression variables were then ranked using a similar approach as described above. To compare my original gene expression variables with the variance-matched random genes, the ranks across all models were averaged (Figure 3.1D), and p-values were obtained from one-tailed Mann-Whitney non-parametric test between the two groups from Elastic Net and Random Forest. To further test my gene expression variables against other gene families, 5 popular gene sets were considered: Receptor tyrosine kinases (RTK), Receptor serine kinases (RSK), Toll like receptors (TLR), MAPK signaling and WNT signaling families. The list of genes in these families were obtained from the HUGO Gene Nomenclature Committee (HGNC)(http://www.genenames.org/cgi-bin/genefamilies/). The same approach as described above for randomly selected genes was used to compare these gene sets with my original gene expression variables (Figure A2.S3)

3.5.13 Distance to nearest gene transcriptional start site (TSS)

Selected 10 kb regions were converted to genomic range objects using the R package "GenomicRanges" (Lawrence et al., 2013). The distance to single nearest gene's transcription start site (TSS) was found using Genomic Regions Enrichment of Annotations Tools (GREAT) (McLean et al., 2010). Genomic regions are associated with nearby genes by first assigning a regulatory domain to every gene in the genome, and then finding genes whose regulatory domains overlap with a given genomic region. I set the association rule parameter in GREAT to "Single nearest gene" with a maximum extension of 1000 kb for definition of regulatory domains. Density plots of distance to TSS are depicted in Supplementary Fig. 2b. The same approach was used for annotating peaks obtained from Figure 3.3 (density plots shown in Figure A2.S5A,C). To obtain the distribution of Illumina probe densities around the TSS, I randomly selected 10000 probes across the arrays and applied the above-described approach to measure the distance to nearest gene's TSS for each probe. Density plots were obtained for the purpose of comparison with the distribution of metabolically regulated peaks (Figure A2.S5B,D).

3.5.14 Identification of metabolically regulated genomic regions

To find peaks of strong association between the met cycle and DNA methylation, I designed a novel scanning method by applying the idea of Manhattan plots from e-QTL analyses to DNA methylation data. In each cancer type, I first selected one of the major enzymes in the met cycle with the highest overall Spearman

correlation with global and local DNA methylations (*BHMT2* in brain, breast, prostate, and liver; *MAT2B* in lung and bladder; and *AHCY* in colon and kidney cancers), and calculated the Spearman correlation between its expression and the beta value of each individual probe across the genome. I then sorted the probes according to genomic coordinates and aligned the –log10 of the p-values obtained from the Spearman correlations along the chromosomes. Next, I applied a sliding window scan for regions of strong association across the genome separately in each cancer type (Figure 3.3A). For this, probes with the highest correlations (top 10% across the genome) were located and a 6 kb window (+3 kb and -3 kb) flanking the genomic coordinate of the original probe was scanned. A region was reported as a "peak" if the following criteria were met: 1- Region included at least 3 probes with a correlation in the same direction as the original probe (positive or negative); 2- At least 80% of all probes within the region had a significant ($p<0.00001$) correlations with met cycle expression. After applying these filters, the selected regions were annotated and genes overlapping with each of the peaks were used for subsequent pathway enrichment analyses. Given the window size and the above criteria, the majority of the identified peaks only overlapped with one unique.

To assess potential bias toward highly methylated regions in the identified regions where correlation of methylation with met cycle expression peaks, I tested 2000 randomly selected probes across the genome. I then evaluated the association between methylation of each probe with the value of its Spearman correlation rho with met cycle expression— I used AHCY in colon cancer as an example in this test (Figure 3.3B).

Finally, an additional filter was applied to rank the identified peaks according to peak shape. For this, the aligned correlation coefficients in each region were assessed with respect to whether they formed a peak according to an information theory score calculated by the R function "turnpoints" (refer to R package "pastecs" (Grosjean and Ibanez, 2014)). This function finds all turning points (peaks and pits) in a series of points (in this case, aligned correlation coefficients), and calculates the information quantity of each turning point using Kendall's information theory. Finally, it measures a p-value against a random distribution of the turning points in a given series, with smaller p-values corresponding to less random shape and a higher probability of a turning point corresponding to a real peak or pit. I selected regions containing turning points with the most significant p-values (lowest 20%) in each cancer type and subsequently tested them for specificity for the met cycle as described in the following section.

3.5.15 Test of specificity of peaks for the met cycle

Each of the selected peaks was tested for specificity of their correlations with the met cycle expression (vs. gene expression in general). For this, 500 genes were randomly selected from the genome in each cancer type, and the Spearman correlation coefficient was measured between their expression and the methylation of every probe within a given peak. The fraction of significant correlations was calculated for all of the 500 genes as well as for the met cycle gene. A randomization q-value was calculated for the met cycle gene by comparing it to the distribution of the correlations

calculated for the 500 random genes. This procedure was repeated separately for each peak in each cancer type and the results are summarized in Figure A2.S8A,B.

3.5.16 Pathway enrichment analyses

Peaks were annotated according to Illumina information and UCSC Ref gene names for genes overlapping with the identified peaks were extracted. Pathway enrichment analysis was performed on the resulting gene list for each cancer type using Enrichr (Chen et al., 2013). Combined scores from Enrichr were used to rank pathways. The Combined score "*c*" is defined as $c=log(p)*z$ where p refers to the p-value from the Fisher's exact test and z is the z-score indicating the deviation from the expected rank. Enrichr first calculates Fisher's exact p-values for many random gene sets to generate a distribution of expected p-values for each pathway in their pathway library. The z-score for deviation from this expected rank is therefore an alternative ranking score and the combined score is considered a corrected form of the enrichment score and p-value, which I used to sort pathways in Figure 3.3C-F and Figure A2.S7A-D. All gene sets in Figure 3.3C-F and Figure A2.S7A-D had Fisher's exact p-values <0.05, and the most highly enriched sets are shown ranked by the combined enrichment scores. Gene set names used in Figure 3.3C-F and Figure A2.S7 follow the convention used and described by Enrichr (*http://amp.pharm.mssm.edu/Enrichr/#stats*). Briefly, gene ontology (GO) sets are shown by GO numbers in parenthesis following their name, epigenetic modifications from the ENCODE histone modifications 2015 project are shown by "-hg19" following gene set names to be distinguished from those from the Epigenomics

Roadmap project, gene sets from the Cancer Cell line Encyclopedia are shown by cell line names following cancer type in upper case, disease signatures from the gene expression omnibus (GEO) are shown in upper case followed by GSE accession numbers, KEGG 2015 and the Human Gene Atlas gene sets are shown in lower case. Refer to Enrichr for a complete list of all gene sets included in more than 70 libraries.

3.5.17 Cancer genes

A list of 12 cancer drivers common in multiple human cancers was considered (Tamborero et al., 2013) (tumor protein p53 (*TP53*), phosphate and tensin homolog (*PTEN*), neuroblastoma RAS viral oncogene homolog (*NRAS*), epidermal growth factor receptor (*EGFR*), isocitrate dehydrogenase 1(*IDH1*), isocitrate dehydrogenase 2 (*IDH2*), CCCTC-binding factor (*CTCF*), von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase (*VHL*), catenin beta 1 (*CTNNB1*), nuclear factor erythroid-2 like 2 (*NFE2L2*), phosphoinositide-3-kinase, regulatory subunit 1 (*PIK3R1*), and ms-related tyrosine kinase 3 (*FLT3*)). These genes were consistently identified as candidate cancer drivers by 4 independent positive selection detection algorithms in a comprehensive pan-cancer analysis of thousands of TCGA tumors (Tamborero et al., 2013). I added to this list, well-known cancer drivers not included in the above list (Kirsten rat sarcoma viral oncogene homolog (*KRAS*), B-Raf proto-oncogene, serine/threonine kinase (*BRAF*), phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha (*PIK3CA*), and breast cancer 1, early onset (*BRCA1*)). In addition to these common cancer drivers, I also considered a number of cancer type-specific genes including receptors important in specific subtypes of cancers (estrogen

receptor 1(*ESR1*), androgen receptor (*AR*), erb-b2 receptor tyrosine kinase 2 (*ERBB2*)). Finally, cancer genes frequently aberrantly methylated in human cancers were also considered (Heyn and Esteller, 2012) (RAS association domain family member-1 (*RASSF1*), glutathione S-transferase pi 1 (*GSTP1*), adenomatous polyposis coli (*APC*), and O-6-methylguanine-DNA methyltransferase (*MGMT*)), together constructing a list of 23 cancer genes for detailed analysis of DNA methylation shown in Figure 3.4.

3.5.18 Evaluation of model performance using randomized responses

In order to test the reliability of the variable contribution results obtained from my gene-specific DNA methylation models, I built two different randomized data sets as control responses, each with the same dimensions as the original response dataset (i.e. the cancer gene DNA methylations). In the first case, I permuted the DNA methylation values of each cancer gene, and repeated the modeling using the met cycle variables. In the second case, I generated random beta-values (from uniform distribution in the range of 0-1) and used those as the response variables in the calculations. I then compared average met cycle variable importance (Random Forest) and variable usage (Elastic Net) from prediction of true cancer gene methylations vs. permuted methylations and randomly generated responses. The Kolmogrov-Smirnov test p-values were calculated between the distributions as illustrated in Figure A2.S9B.

3.5.19 Evaluation of model performance using randomized predictors

Using prostate cancer as an example, Lucas Mentch performed simulation tests to determine whether the Random Forest as a methodology is able to utilize the information in the prediction variables beyond what could be expected if the predictors were only random noise and unrelated to the response. To investigate this, he modeled methylation in the prostate cancer dataset at 3 example cancer loci (*GSTP1, RASSF1, and PITX2*). These genes were selected based on previous evidence indicating the critical importance of their aberrant methylation in prostate cancer (Heyn and Esteller, 2012; Litovkin et al., 2014). As controls, he generated 3 additional datasets. For the first dataset, he copied the exact response as the *GSTP1* methylation, but randomly generated a predictor variable set of the same dimensions as the original variable set by sampling from a standard normal distribution. That is, each observation on each variable is a sample from a normal distribution of unit variance and should therefore have no relationship to the response. The other two datasets were generated in the same fashion, using *RASSF1* and *PITX2* methylation as responses and randomly generated variable sets as predictors. To assess the performance of the Random Forest computations, Lucas Mentch compared the mean squared error (MSE) from predictions made using the original data with those made by the datasets consisting of random variables unrelated to the responses. For each of the three responses, he randomly divided the data into training and test sets, generated a total of 100 simulations consisting of 500 decision trees, and compared the resulting MSEs of the predictions made on the test points. Results are summarized in Figure A2.S10A-D.

To quantify the improvement in the Random Forest algorithm by using the original variables over the randomly simulated variables, Lucas Mentch defined an

improvement metric (MSE-Imp), describing the relative improvement in prediction accuracy:

$$\text{MSE-Imp} := \frac{\mathbf{MSE-rand}}{\mathbf{MSE-orig}}$$

where MSE-rand is the average MSE calculated using the random simulated variables and MSE-orig is the average MSE calculated using the original variables.

In this test, another simulated dataset of the same dimensions as the original dataset was generated where the variables and response were linearly related via the following equation:

$$Y = \sum_{i=1}^{P} \boldsymbol{\beta}_i \boldsymbol{X}_i + \boldsymbol{\varepsilon}$$

To generate this linear dataset, he sampled the value of each prediction variable Xi from a standard normal distribution and the noise ε from a normal distribution with mean 0 and standard deviation 0.05. The values of the coefficients βi were selected uniformly at random from the interval [0,1]. He then measured the MSE improvement (MSE-Imp) for the linear dataset using the same approach as MSE improvements for the original datasets were calculated (explained in the previous paragraph). This allowed us to compare a linearly simulated dataset to my real dataset. Results are shown in Figure A2.S10E-F.

3.5.20 Network construction

Genes in the serine, glycine, one-carbon (SGOC) network (including the met cycle genes) that were among the most highly ranked variables (top 15%) in at least 4 of the cancer datasets according to the Elastic Net models and at least 3 of the cancer datasets according to the Random Forest models were selected. A metabolic network consisting of these enzymes was then constructed using MetScape (Gao et al., 2010) where nodes represent genes and metabolites, and edges represent biochemical links. I fixed the node size for metabolites but adjusted node sizes for genes to correspond to the number of cancers in which each variable was highly ranked (among the top 15% of all variables) (Figure 3.4G). For nodes not directly connected to the rest of the network, I manually added dashed lines where appropriate.

3.5.21 Survival analyses

In each cancer type, the average error of prediction of DNA methylation at cancer loci was measured for each patient across all Elastic Net models using only met cycle variables for prediction. Patients were then divided into 2 groups based on predictability of their methylation by the met cycle activity ("predictable"= below-median prediction error, "not predictable"= above-median prediction error). To estimate overall survival time, "days-to-death" was used with vital status information and last follow-up date used to right-censor subjects (subjects alive at last follow-up were censored from the analysis beyond their last follow up date). The relationship between survival and predictability was then analyzed using the "survfit" function in the R package "survival" (Therneau, 2015) and visualized by Kaplan-Meier curves. Log rank test p-values were calculated by fitting models of overall survival to patients'

"predictability" group assignments using the "survdiff" function in the survival package for each cancer type separately. Results are depicted in Figure 3.5.

3.5.22 Multivariate Cox regression

In the three cancer types (brain, liver, and kidney) where univariate analysis showed a highly significant difference in survival between the predictable and non-predictable groups as described above, and also the sample size allowed for sufficient power to perform multivariate analysis, I used relevant clinical and mutational factors as covariates and repeated the survival analysis. The following factors were individually tested as covariates in separate models of overall survival along with "predictability" status as the fixed effect: Brian cancer: all frequent somatic mutations, histological diagnosis, age, gender, and initial weight; Liver cancer: all frequent somatic mutations, tumor stage, history of other malignancies, and residual tumor; Kidney cancer: all frequent somatic mutations, age, and race. In each case, the results of regression using the "coxph()" function in R provided the p-value for the significance of the predictability status when modeling overall survival in the presence of covariates.

3.5.23 Software

All computational and statistical analyses were done using R 3.1.2 (R-Core-Team, 2014). Distribution plots, box-plots, scatter-plots, and bar-plots were made in GraphPad Prism version 6 (GraphPad Software, San Diego California USA,

www.graphpad.com). Circular plots were generated using Circos (Krzywinski et al., 2009).

3.5.23 Code availability

R script is available through the following Github repository (https://github.com/mahyam/DNA-methylation-and-metabolism-R-code).

3.5.24 Data availability

All data used in this study was obtained from the TCGA data portal available online at: https://gdc.nci.nih.gov/.

3.6. References

Anderson, O.S., Sant, K.E., and Dolinoy, D.C. (2012). Nutrition and epigenetics: an interplay of dietary methyl donors, one-carbon metabolism and DNA methylation. The Journal of nutritional biochemistry *23*, 853-859.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P., van Dijk, C.M., Tollenaar, R.A., et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nature genetics *44*, 40-46.

Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32.

Brosnan, J.T., da Silva, R.P., and Brosnan, M.E. (2011). The metabolic burden of creatine synthesis. Amino acids *40*, 1325-1331.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cancer Genome Atlas Research, N. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature *499*, 43-49.

Cancer Genome Atlas Research, N. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315-322.

Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics *45*, 1113-1120.

Carone, D.M., and Lawrence, J.B. (2013). Heterochromatin instability in cancer: from the Barr body to satellites and the nuclear periphery. Seminars in cancer biology *23*, 99-108.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nature reviews. Genetics *10*, 295-304.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery *2*, 401-404.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics *14*, 128.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Costello, J.C., Heiser, L.M., Georgii, E., Gonen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. Nature biotechnology *32*, 1202-1212.

Duncan, C.G., Barwick, B.G., Jin, G., Rago, C., Kapoor-Vazirani, P., Powell, D.R., Chi, J.T., Bigner, D.D., Vertino, P.M., and Yan, H. (2012). A heterozygous IDH1R132H/WT mutation induces genome-wide alterations in DNA methylation. Genome research *22*, 2339-2355.

Ehrlich, M., and Lacey, M. (2013). DNA hypomethylation and hemimethylation in cancer. Advances in experimental medicine and biology *754*, 31-56.

Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schubeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS genetics *9*, e1003994.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software *33*, 1-22.

Gaidatzis, D., Burger, L., Murr, R., Lerch, A., Dessus-Babus, S., Schubeler, D., and Stadler, M.B. (2014). DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. PLoS genetics *10*, e1004143.
Gao, J., Tarcea, V.G., Karnovsky, A., Mirel, B.R., Weymouth, T.E., Beecher, C.W., Cavalcoli, J.D., Athey, B.D., Omenn, G.S., Burant, C.F., et al. (2010). Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics *26*, 971-973.
Gentles, A.J., Newman, A.M., Liu, C.L., Bratman, S.V., Feng, W., Kim, D., Nair, V.S., Xu, Y., Khuong, A., Hoang, C.D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. Nature medicine *21*, 938-945.
Grosjean, P., and Ibanez, F. (2014). pastecs: Package for Analysis of Space-Time Ecological Series. R package version 1.3-18.
Gut, P., and Verdin, E. (2013). The nexus of chromatin regulation and intermediary metabolism. Nature *502*, 489-498.
Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nature genetics *43*, 768-775.
Heyn, H., and Esteller, M. (2012). DNA methylation profiling in the clinic: applications and challenges. Nature reviews. Genetics *13*, 679-692.
Ho, V., Ashbury, J.E., Taylor, S., Vanner, S., and King, W.D. (2015). Gene-specific DNA methylation of DNMT3B and MTHFR and colorectal adenoma risk. Mutation research *782*, 1-6.
Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome research *22*, 246-258.
Jia, L., Li, J., He, B., Jia, Y., Niu, Y., Wang, C., and Zhao, R. (2016). Abnormally activated one-carbon metabolic pathway is associated with mtDNA hypermethylation and mitochondrial malfunction in the oocytes of polycystic gilt ovaries. Scientific reports *6*, 19436.
Johnston, S.J., and Carroll, J.S. (2015). Transcription factors and chromatin proteins as therapeutic targets in cancer. Biochimica et biophysica acta *1855*, 183-192.
Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews. Genetics *13*, 484-492.
Jung, M., and Pfeifer, G.P. (2015). Aging and DNA methylation. BMC biology *13*, 7.
Kaelin, W.G., Jr., and McKnight, S.L. (2013). Influence of metabolism on epigenetics and disease. Cell *153*, 56-69.
Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of American Statistical Association *53*, 457-481.
Katoh, Y., Ikura, T., Hoshikawa, Y., Tashiro, S., Ito, T., Ohta, M., Kera, Y., Noda, T., and Igarashi, K. (2011). Methionine adenosyltransferase II serves as a transcriptional corepressor of Maf oncoprotein. Molecular cell *41*, 554-566.
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome research *19*, 1639-1645.
Landau, D.A., Clement, K., Ziller, M.J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., et al. (2014). Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. Cancer cell *26*, 813-825.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS computational biology *9*, e1003118.
Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18-22.
Litovkin, K., Joniau, S., Lerut, E., Laenen, A., Gevaert, O., Spahn, M., Kneitz, B., Isebaert, S., Haustermans, K., Beullens, M., et al. (2014). Methylation of PITX2, HOXD3, RASSF1 and TDRD1 predicts biochemical recurrence in high-risk prostate cancer. Journal of cancer research and clinical oncology *140*, 1849-1861.
Locasale, J.W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. Nature reviews. Cancer *13*, 572-583.
Lokk, K., Modhukur, V., Rajashekar, B., Martens, K., Magi, R., Kolde, R., Koltsina, M., Nilsson, T.K., Vilo, J., Salumets, A., et al. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome biology *15*, r54.
Mack, S.C., Witt, H., Piro, R.M., Gu, L., Zuyderduyn, S., Stutz, A.M., Wang, X., Gallo, M., Garzia, L., Zayne, K., et al. (2014). Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. Nature *506*, 445-450.
Marchal, C., and Miotto, B. (2015). Emerging concept in DNA methylation: role of transcription factors in shaping DNA methylation patterns. Journal of cellular physiology *230*, 743-751.
McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.
Mehrmohamadi, M., Liu, X., Shestov, A.A., and Locasale, J.W. (2014). Characterization of the usage of the serine metabolic network in human cancer. Cell reports *9*, 1507-1519.
Mentch, L., and Hooker, G. (2015). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. The Journal of Machine Learning Research *in press*.
Mentch, S.J., Mehrmohamadi, M., Huang, L., Liu, X., Gupta, D., Mattocks, D., Gomez Padilla, P., Ables, G., Bamman, M.M., Thalacker-Mercer, A.E., et al. (2015). Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. Cell metabolism *22*, 861-873.
Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology *12*, R41.
Nakayama, T., Watanabe, M., Suzuki, H., Toyota, M., Sekita, N., Hirokawa, Y., Mizokami, A., Ito, H., Yatani, R., and Shiraishi, T. (2000). Epigenetic regulation of androgen receptor gene expression in human prostate cancers. Laboratory investigation; a journal of technical methods and pathology *80*, 1789-1796.
Pfalzer, A.C., Choi, S.W., Tammen, S.A., Park, L.K., Bottiglieri, T., Parnell, L.D., and Lamon-Fava, S. (2014). S-adenosylmethionine mediates inhibition of inflammatory response and changes in DNA methylation in human macrophages. Physiological genomics *46*, 617-623.
R-Core-Team (2014). R: A Language and Environment for Statistical Computing.
Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.
Robertson, K.D. (2001). DNA methylation, methyltransferases, and cancer. Oncogene *20*, 3139-3155.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., and Karp, P.D. (2005). Computational prediction of human metabolic pathways from the complete human genome. Genome biology *6*, R2.
Sahar, S., and Sassone-Corsi, P. (2009). Metabolism and cancer: the circadian clock connection. Nature reviews. Cancer *9*, 886-896.
Schubeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321-326.
Sproul, D., Nestor, C., Culley, J., Dickson, J.H., Dixon, J.M., Harrison, D.J., Meehan, R.R., Sims, A.H., and Ramsahoye, B.H. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. Proceedings of the National Academy of Sciences of the United States of America *108*, 4364-4369.
Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490-495.
Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific reports *3*, 2650.
Therneau, T. (2015). A Package for Survival Analysis in S_. version 2.38.
Timp, W., and Feinberg, A.P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nature reviews. Cancer *13*, 497-510.
Turcan, S., Rohle, D., Goenka, A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W., Lu, C., Ward, P.S., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature *483*, 479-483.
van Dongen, J., Nivard, M.G., Willemsen, G., Hottenga, J.J., Helmer, Q., Dolan, C.V., Ehli, E.A., Davies, G.E., van Iterson, M., Breeze, C.E., et al. (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. Nature communications *7*, 11115.
Waldmann, P., Meszaros, G., Gredler, B., Fuerst, C., and Solkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers in genetics *4*, 270.
Yang, M., and Park, J.Y. (2012). DNA methylation in promoter region as biomarkers in prostate cancer. Methods in molecular biology *863*, 67-109.
Yang, X., Han, H., De Carvalho, D.D., Lay, F.D., Jones, P.A., and Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer cell *26*, 577-590.
Zhu, Y., Qiu, P., and Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. Nature methods *11*, 599-600.
Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477-481.
Zou, H., and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society *Series B*, 301–320.

CHAPTER 4: INVESTIGATING THE DETERMINANTS OF METHIONINE IN THE HUMAN SERUM [4]

4.1 Abstract

Methylation of DNA and histones are two of the major epigenetic mechanisms known to regulate gene expression. A direct link between epigenetics and the activity of the methionine cycle as the source of the methyl donor for methylation reactions has previously been established. In order to assess the validity of this relationship in humans for potential dietary interventions, it is critical to quantitatively assess inter-individual variation in methionine levels. In this work, I design a computational model to characterize variation in serum methionine in healthy human subjects and identify its determinants. I show that methionine variability in fasting serum is commensurate with concentrations needed for epigenetic changes and can be partly explained by diet and clinical factors.

4.2 Introduction

The one-carbon (1-C) metabolic network is considered the cellular source of S-adenosyl-methionine (SAM) — the universal methyl donor in all methylation reactions (Crider et al., 2012). Methionine is converted to SAM through the action of

[4] Some of the text was published in: Mentch S.J., Mehrmohamadi M, Huang L., Liu X., Gupta D., Mattocks D., Gómez Padilla P., Ables G., Bamman M.M., Thalacker-Mercer A.E., Nichenametla S. , Locasale J.W., Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. *Cell Metabolism* 22, 861-873 (2015).

methionine adenosyl transferase (MAT) enzymes and this provides an important regulatory link between 1-C metabolism and epigenetics (Locasale, 2013). Studies in human pluripotent stem cells have demonstrated that a depletion of methionine, which is the precursor to SAM could lead to changes in histone and DNA methylation (Shiraki et al., 2014). However, these changes are also accompanied by widespread induction of stress response pathways and cell death confounding the interpretation of whether the epigenetic changes occurred directly through the sensing of SAM status. Furthermore, previous studies have provided evidence that under pathological conditions, aberrant expression of nicotinamide N-methyltransferase (NNMT)—an enzyme that metabolizes SAM— has profound biological consequences resulting from changes in histone methylation (Kraus et al., 2014; Ulanovskaya et al., 2013).

Previous work from our group has provided evidence for the regulatory role of 1-C metabolism in epigenetics. We have reported a significant contribution from expression levels of methionine cycle enzymes in predicting DNA methylation status in human tumors (Mehrmohamadi et al., 2016). Furthermore, we have performed methionine restriction experiments in cell lines as well as in mice to assess the link between the methionine cycle and epigenetics (Mentch et al., 2015). Mentch et al. found that both SAM levels and the SAM/SAH ratio can be quantitatively tuned through changes in the metabolic flux of the methionine cycle to affect histone methylations to control numerous physiological processes including direct feedback regulation for the maintenance of homeostasis in 1-C metabolism and the activity of genes involved in cancer and cell fate (Mentch et al., 2015). She further illustrated this link by dietary restriction of methionine in mice and reported alterations in histone

methylation in the liver (Mentch et al., 2015). However, whether 1-C metabolism's relationship with methylation also occurs in humans at physiological conditions requires further investigation.

Here, I use computational modeling and integrate clinical and environmental information including diet records to determine the contribution of different factors in explaining variation in human serum methionine levels. I identify specific dietary sources that may have a significant impact on serum methionine levels. Together, my results complement Mentch's findings and confirm that the regulation of epigenetics through 1-C metabolism occurs at physiologically relevant concentrations of methionine, and diet explains a significant portion of overall variation in serum methionine levels. The results of this work have been published as a separate section in the same manuscript (Mentch et al., 2015).

4.3 Results

4.3.1 Humans exhibit variability in methionine levels

I asked whether variability in methionine metabolism exists in humans and how it can be regulated. Standard clinical parameters and a record of dietary intake over four days was considered to reflect variations in habitual diet as is standard practice in clinical nutrition (Levine et al., 2014) across a cohort of healthy human subjects. Fasting serum was collected and subjected to a metabolomics analysis (Figure 4.1A). We then performed an unsupervised hierarchical clustering of the

nutrient intake for each subject and found sets of defined modules that were able to classify the subjects into discrete dietary behaviors such as groups that are high in fruits and vegetables or carbohydrates (Figure 4.1B). We next measured the concentration of methionine along with a panel of amino acids in these subjects (Figure 4.1C). Strikingly, the concentration of methionine exhibits substantial variation with values ranging from 3 to 30 μM, and of all amino acids, methionine exhibited the largest variation (Figure 4.1.C). This variation in concentration is on the same order as that needed to induce changes in histone methylation in cells and in mice (Mentch et al., 2015).

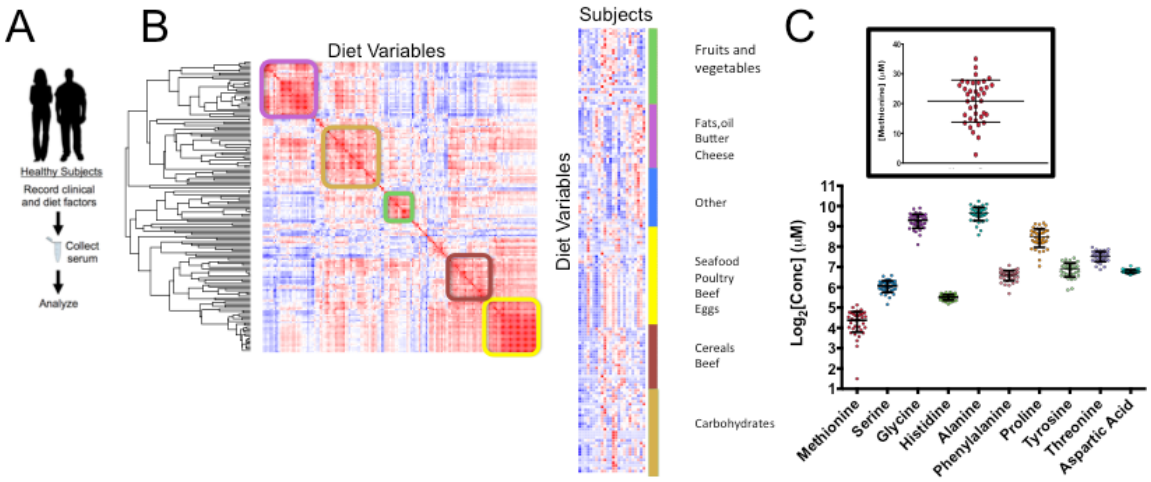


Figure 4.1 – Methionine and metabolic variation in human subjects.[5]

A) Measurement of serum methionine and clinical and dietary variables in human subjects.

B) (left) Hierarchical clustering of the distance matrix diet variables. (right) k-means clustering of subjects and diet variables (N=24).

C) Absolute concentrations of amino acids in fasting serum in 38 human subjects.

[5] Credit goes to Samantha J. Mentch: Mentch, S.J., Mehrmohamadi, M., Huang, L., Liu, X., Gupta, D., Mattocks, D., Gomez Padilla, P., Ables, G., Bamman, M.M., Thalacker-Mercer, A.E., et al. (2015). Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. Cell metabolism *22*, 861-873.

We next considered the correlations among various serum metabolites that were quantified (Figure 4.2A), and observed that levels of metabolites in the taurine and glutamine metabolic pathways correlate with the methionine metabolism intermediates in the serum (Figure 4.2B). In addition, methionine in the serum correlated with N,N,N-trimethyllysine and N-methylglycine (sarcosine), both of which are methylated by the transfer of methyl groups from SAM, suggesting that methionine levels in the serum are indicative of cellular methylation status (Figure 4.3A) with clear patterns of food consumption that corresponded to both high and low methionine intake (Figure 4.3B).

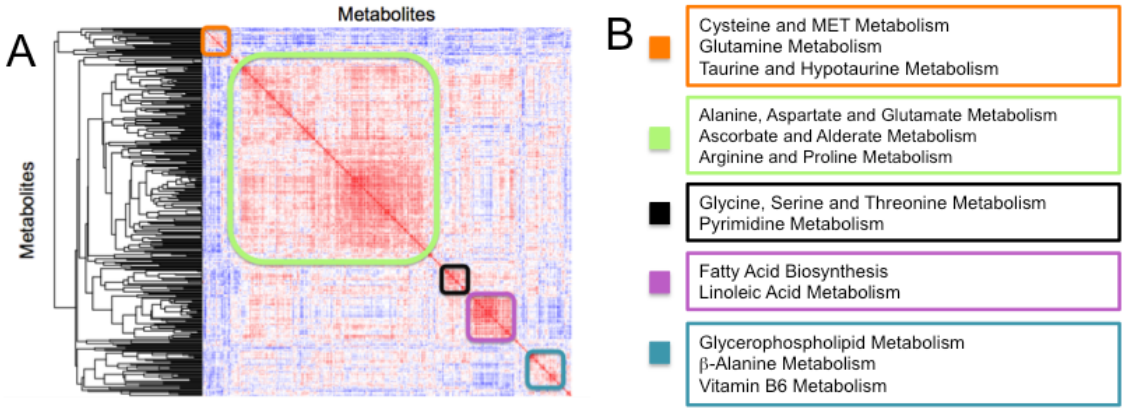


Figure 4.2— Serum metabolic profiles in human subjects. [6]

A) Measurement of a metabolomics profile across the human cohort (N=38). Results of unsupervised clustering of a distance matrix for metabolites are shown.

B) Pathways that correspond to clusters observed in (A). Pathways were identified from consideration of the highest pathway impacts of all metabolites contained in the specified cluster denoted by different colored boxes.

[6] Credit goes to Samantha J. Mentch.

Vegetable-based nutrients such as fiber correlated with low methionine levels and age, body weight and fat intake correlated with high methionine levels. Surprisingly, protein intake exhibited no correlation with methionine. An analysis of the metabolites and pathways that correlated with methionine levels revealed pathways related to ketogenesis and amino metabolism (Figure 4.3C). Taken together, these findings demonstrate that the variability in methionine concentration in humans is on the same scale as that needed to induce alterations in histone methylation and that these differences correlate with changes in diet and health status.

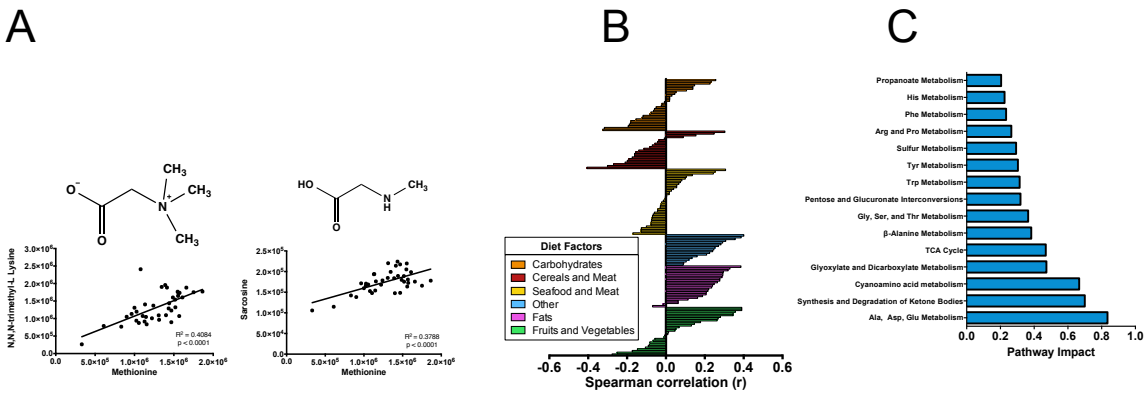


Figure 4.3— Analysis of methionine-correlated serum metabolites. [7]

A) Correlation of methionine in the serum with methylated serum metabolites, N,N,N-trimethyllysine and sarcosine.

B) Correlation of methionine concentrations with dietary variables obtained from habitual diet records.

C) Correlation of methionine concentrations with fasting serum metabolite levels.

4.3.2 A computational model identifies factors that determine methionine levels

[7] Credit goes to Samantha J. Mentch.

Having identified associations of methionine with diet and other factors, I next built a computational model to identify the direct influences on methionine concentration. I considered a mixed effects model that aims to identify causal features in high dimensional data and thus identifies factors that give rise to the variation in methionine levels (see Methods). I considered dietary intake variables, clinical variables such as age, gender, and body composition measured by Dual X-ray Absorptiometry (DEXA). To reduce the dimensionality of the data matrix, I filtered the variables first according to their correlations with methionine, and then carried out a principle components analysis on the DEXA variables (Figure 4.4A), and finally checked the resulting twelve variables for collinearity (Figure 4.4B). I then carried out a regression with maximum likelihood estimation (Methods) using least squares to obtain a model with good fit ($P < 10^{-4}$, F test) to the experimentally measured methionine concentrations (Figure 4.5A,B).

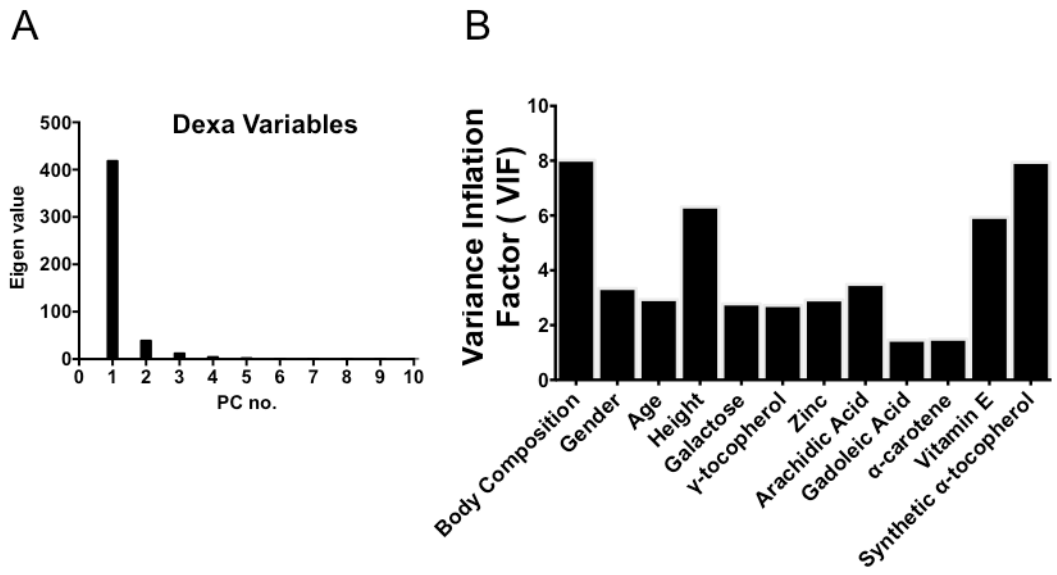


Figure 4.4— A computational model identifies determinants of methionine variability.

A) Eigenvalue spectrum form principal components analysis of DEXA variables.

B) Calculation of variance inflation factor for each variable.

An analysis of the coefficients revealed several contributions to methionine variation including age, body composition and gender (with maleness contributing to a positive influence) and diet including variables known to be associated with methionine such as zinc and tocepherol (Figure 4.5C). Using established guidelines (Methods), we was found that the diet variables could be related to major sources of food intake (Figure 4.5D) with for example, fats, seafood and meat contributing to higher concentrations of methionine.

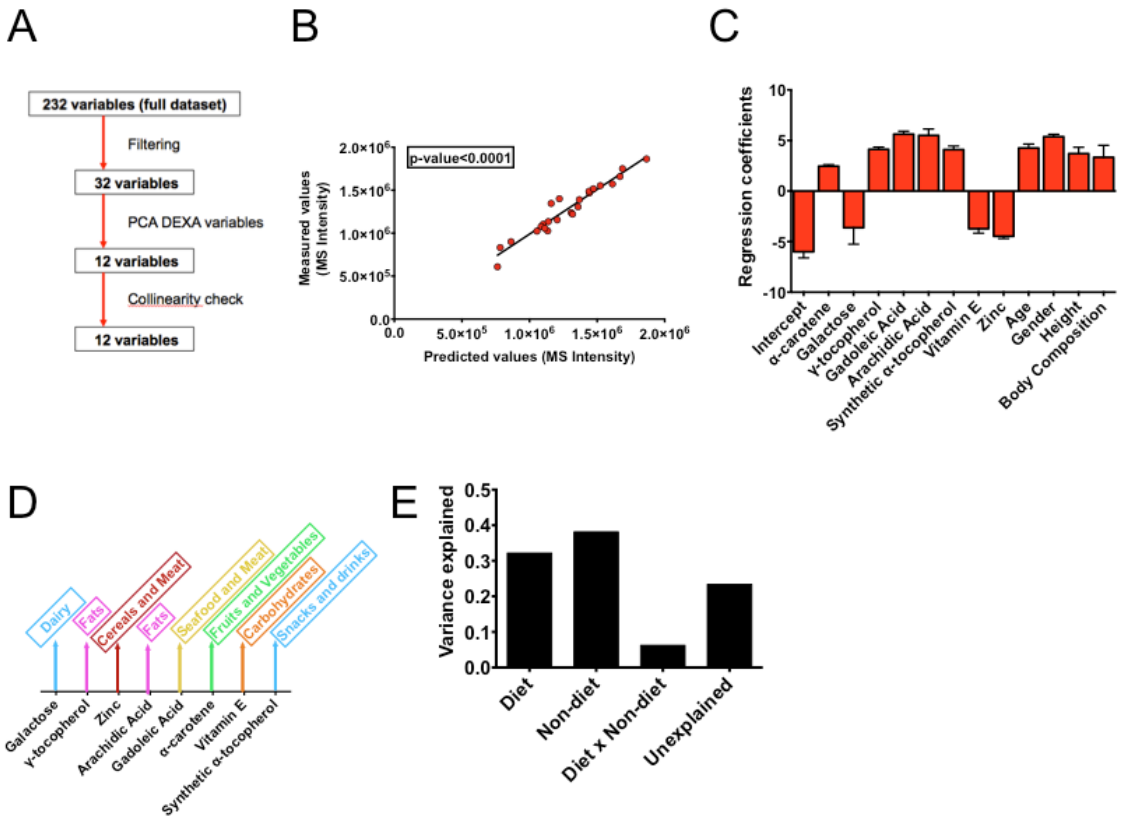


Figure 4.5— A computational model identifies determinants of methionine variability.

A) Overview of variable selection for the computational model.

B) Predicted versus measured methionine levels in human subjects (model fit).

C) Regression coefficients. Error bars are obtained from maximum likelihood estimates.

D) Schematic of dietary factors that contribute to each modeled variable.

E) Results from variance partitioning.

4.3.3 Variance partitioning quantifies relative contributions of factors explaining methionine variation

Finally, my collaborator Lei Huang and I calculated the proportion of variance explained by each of the categories of variables in the study using variance

partitioning (Methods). We found that about thirty percent of the variation in human serum methionine is explained by diet, about thirty percent explained by clinical variables including gender and age and the remaining unaccounted variance is likely due to genetic differences (Figure 4.5E). These results indicate that physiological methionine concentrations observed in humans are determined by several factors including diet, suggesting potential for dietary intervention in cases of pathological states where methionine is implicated to play a role.

4.4 Discussion

Previous *in silico*, *in vitro*, and *in vivo* experiments have provided evidence for the contribution of 1-C metabolism to cellular epigenetics through the activity of the methionine cycle. The goal of the present work was to assess the relevance of these findings in humans under physiological concentrations of methionine. The variation of methionine in the serum of healthy individuals was found to be largest across a panel of serum amino acids. Although these measurements are by no means exhaustive of human population dynamics, concentrations in many individuals in our cohort were found to be far lower than the concentration required for inducing changes in methylation levels. Computational modeling of diet and clinical variables found that about thirty percent of the variation could be due to fundamental clinical variables such as age, body composition, and gender, and about thirty percent was explained by variations in individuals' diets. Approximately 20% of the variation remained unexplained by our models, which we speculate to be at least in part due to genetic factors. In the future, the results from these models could further be integrated with

genomics data to specifically define the genetic contributions as have been identified to associate with serum metabolite levels (Suhre et al., 2011). Nevertheless it is tempting to speculate that the intake of basic dietary factors such as vegetables and fat could mediate human epigenetics through modulation of methionine metabolism. Previous studies have demonstrated a wide range of physiological responses to dietary methionine restriction, the most well known example of which is extension of lifespan (Ables et al., 2016). However, detailed molecular and cellular mechanisms through which methionine restriction acts remain largely unknown. Our results suggest potentials for epigenetic alterations through dietary intervention experiments using varying levels of methionine in both normal and pathological conditions.

4.5 Methods

4.5.1 Human Subjects

Serum samples, 4-day diet records, and body composition results (via DEXA scan) on 24 de-identified healthy older adults were provided by Marcas Bamman at the University of Alabama at Birmingham (UAB). As part of the UAB Institutional Review Board-approved parent project, all 24 subjects agreed to have their samples and data used for future research.

4.5.2 Clinical Nutrition Studies

A four-day diet record was collected according to previous standards of recording diet to reflect habitual behaviors as has been previously described (Levine et al., 2014). Surveys were converted into nutritional variables according to previously described procedures using standard software (Levine et al., 2014). Nutrition data system for research (NDSR) dietary analysis software (University of Minnesota), a comprehensive food and nutrient database, was used to determine the average macro- and micronutrients consumed over the four-day period as previously described (Thalacker-Mercer et al., 2009). Before serum was collected, each subject was subjected to overnight fasting. Metabolites from serum were extracted using a previously described protocol(Shestov et al., 2014). All additional clinical variables were recorded according to previously described methods(Thalacker-Mercer et al., 2013; Thalacker-Mercer et al., 2009).

4.5.3 Metabolite Extraction

For culture from adherent cells, the media was quickly aspirated and cells were washed with cold PBS on dry ice. Then, 1mL of extraction solvent (80% methanol/water) cooled to -80°C was added immediately to each well, and the dishes were transferred to -80°C for 15 min. Plates were removed and cells were scraped into the extraction solvent on dry ice. For tissue, the sample was homogenized in liquid nitrogen and then 5 to 10 mg was weighed in a new Eppendorf tube. Ice-cold extraction solvent (250μL) was added to each tissue sample and homogenized using a tissue homogenizer. The homogenate was incubated on ice for 10 min. For plasma or serum, 20 μL was transferred to a new Eppendorf tube containing 80 μL HPLC grade water. Next, 400 μL of ice-cold methanol was added to the sample for a final methanol concentration of 80% (v/v). Samples were incubated on ice for 10 min. All metabolite extracts were centrifuged at 20,000g at 4°C for 10 min. Finally, the solvent in each sample was evaporated in a Speed Vacuum for metabolomics analysis. For polar metabolite analysis, the cell extract was dissolved in 15 μL water and 15 μL methanol/acetonitrile (1:1 v/v) (LC-MS optima grade, Thermo Scientific). Samples were centrifuged at 20,000g for 10min at 4°C and the supernatants were transferred to Liquid Chromatography (LC) vials. The injection volume for polar metabolite analysis was 5 μL.

4.5.4 Liquid Chromatography

An Xbridge amide column (100 x 2.1 mm i.d., 3.5 μm; Waters) is employed on a Dionex (Ultimate 3000 UHPLC) for compound separation at room temperature. The mobile phase A is 20 mM ammonium acetate and 15 mM ammonium hydroxide in water with 3% acetonitrile, pH 9.0 and mobile phase B is acetonitrile. Linear gradient as follows: 0 min, 85% B; 1.5 min, 85% B, 5.5 min, 35% B; 10min, 35% B, 10.5 min, 35% B, 14.5 min, 35% B, 15 min, 85% B, and 20 min, 85% B. The flow rate was 0.15 ml/min from 0 to 10 min and 15 to 20 min, and 0.3 ml/min from 10.5 to 14.5 min. All solvents are LC-MS grade and had been purchased from Fisher Scientific.

4.5.5 Mass Spectrometry

The Q Exactive MS (Thermo Scientific) is equipped with a heated electrospray ionization probe (HESI), and the relevant parameters are as listed: evaporation temperature, 120 °C; sheath gas, 30; auxiliary gas, 10; sweep gas, 3; spray voltage, 3.6 kV for positive mode and 2.5 kV for negative mode. Capillary temperature was set at 320°C, and S-lens was 55. A full scan range from 60 to 900 (m/z) was used. The resolution was set at 70,000. The maximum injection time was 200 ms. Automated gain control (AGC) was targeted at 3,000,000 ions.

4.5.6 Metabolomics and Data Analysis

Raw data collected from LC-Q Exactive MS is processed on Sieve 2.0 (Thermo Scientific). Peak alignment and detection are performed according to the protocol described by Thermo Scientific. For a targeted metabolite analysis, the

method “peak alignment and frame extraction” is applied. An input file of theoretical m/z and detected retention time of 263 known metabolites is used for targeted metabolites analysis with data collected in positive mode, while a separate input file of 197 metabolites is used for negative mode. M/Z width is set at 10 ppm. The output file including detected m/z and relative intensity in different samples is obtained after data processing. Dot plots and other quantitation and statistics were calculated and visualized with the Graphpad prism software package.

4.5.7 Computational Modeling

In brief, a set of 12 predictor variables were obtained after a variable selection process that incorporates methionine correlations, a dimensional reduction of the Dual-energy X-ray Absorptiometry (DEXA) data using principal components analysis, and a collinearity assessment using the variance inflation factor. I next fitted a mixed effects linear model including the set of predictor variables, a random effect term and noise. Model selection was carried out using exhaustive sampling and optimization of the Akaike information criteria. Variance contributions for each variable and random effect were summed to define the total contribution of variance for each variable and unexplained factor. The full details of the modeling are contained in the Supplemental Information.

4.5.7.1 Detailed Description

The dataset consists of 24 samples of serum methionine concentration (SMC) and 233 predictor variables. To understand what factors determine SMC, I develop a linear mixed model (a type of linear regression analysis) following the steps:

The final model:

$$\boldsymbol{y_i = \beta_0 + \sum_k \beta_k X_{i,k} + Z_i + \varepsilon_i}$$

where $\boldsymbol{i = 1, \dots, 24}$ indexes the samples, $\boldsymbol{y_i}$ denotes SMC for the i-th sample, $\boldsymbol{X_{i,k}}$ ($\boldsymbol{k = 1, \dots, 12}$) represent the 12 predictor variables that I model as fixed effect (intercept, alpha carotene, galactose, gamma tocopherol, gadoleic acid, arachidic acid, synthetic alpha tocopherol, vitamin E, zinc, age, gender, height, body composition), $\boldsymbol{Z_i}$ denotes the i-th sample that I model as random effect, and $\boldsymbol{\beta_0, \beta_1, \beta_2, \beta_3}$ and $\boldsymbol{\epsilon_i}$ are the standard parameter and noise terms in linear regression.

4.5.7.2 Variable Selection

Since the number of predictors vastly outnumber the sample size and presumably only a small subset of the predictors are related to SMC, I first have a filtering step to pick out the predictors that significantly correlate with SMC. To do this, each of the 233 predictors is individually fit to the following linear mixed model. A p-value cutoff of 0.05 for the significance of correlation is used, after which 34 variables are left; no correction for multiple testing is conducted in order to include as many potentially relevant variables as possible. I mention that a linear mixed model,

rather than an ordinary linear model, is chosen to account for the fact that some samples are not independent but come from the same individuals.

21 of the 34 variables are Dual-energy X-ray absorptiometry (DEXA) variables that measure bone mineral density. An inspection of the correlation matrix and the principal components analysis (PCA) spectrum revealed that they are highly correlated with each other. Therefore, I condensed each of the 21 variables into a single one by using the first principal component. Large gaps between the leading and following eigenvalues in the PCA spectrum suggests that little information is lost in such a condensation.

I performed a standard procedure of variable selection using the remaining 14 variables to achieve a good balance of model predictive power and simplicity. I follow the suggestion of Faraway and use the Akaike information criterion (AIC) as my criterion for model selection: the model with the lowest AIC is selected. With only 14 candidate variables, I performed an exhaustive evaluation of the AIC of all possible models. This lead to the selection of a model with 12 variables.

Among the final 12 variables, it is important to assess whether there is collinearity. Collinearity refers to the situation in which some predictors contain heavily redundant information and causes problems for parameter identifiability and model interpretation. None of the variables was significantly correlated with any of the others (corresponding to an often-suggested threshold of 10 for variance inflation factor), confirming a lack of collinearity among my 12 variables.

4.5.7.3 Model Analysis

For a multiple linear regression $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, a common interpretation is of a geometric kind: given vector $\boldsymbol{y}$, find a linear combination of the column vectors of $\boldsymbol{X}$, $\hat{\boldsymbol{y}} = \boldsymbol{X\beta}$, such that $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|$ is the smallest, or equivalently, $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to $\hat{\boldsymbol{y}}$. When the set of predictors has a natural partition, such as diet vs. non-diet variables in my case, the model can be written as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X\beta}$ and $\boldsymbol{Z\gamma}$ correspond to the two groups of predictors.

The goodness-of-fit by the linear model is often measured by the so-called *coefficient of determination*, $\boldsymbol{R^2}$, which also has a geometric interpretation: let $\boldsymbol{SST} = \sum_{\boldsymbol{i}}(\boldsymbol{y_i} - \overline{\boldsymbol{y}})^2 = \|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2$ be the total sum of squares (variation), it can be shown that $\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2 = \|\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2 + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \boldsymbol{SSE} + \boldsymbol{SSR}$, where SSE and SSR stand for the explained and residual sum of squares, respectively; $\boldsymbol{R^2}$ is then the proportion of explained sum of squares (variation) out of the total variation, $\frac{\boldsymbol{SSE}}{\boldsymbol{SST}} = \frac{\|\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2}$. When the predictors have two groups, $\hat{\boldsymbol{y}} = \hat{\boldsymbol{x}} + \hat{\boldsymbol{z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$ and $\overline{\boldsymbol{y}} = \overline{\boldsymbol{x}} + \overline{\boldsymbol{z}}$. Hence SSE can be partitioned into three parts: $\boldsymbol{SSE} = \|\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2 = \|(\hat{\boldsymbol{x}} + \hat{\boldsymbol{z}}) - (\overline{\boldsymbol{x}} + \overline{\boldsymbol{z}})\|^2 = \|(\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}}) + (\hat{\boldsymbol{z}} - \overline{\boldsymbol{z}})\|^2 = \|\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}}\|^2 + \|\hat{\boldsymbol{z}} - \overline{\boldsymbol{z}}\|^2 + 2\langle\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}}, \hat{\boldsymbol{z}} - \overline{\boldsymbol{z}}\rangle$; plugging it into the definition of $\boldsymbol{R^2}$, I have

$$\boldsymbol{R^2} = \frac{\boldsymbol{SSE}}{\boldsymbol{SST}} = \frac{\|\hat{\boldsymbol{y}} - \overline{\boldsymbol{y}}\|^2}{\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2} = \frac{\|\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}}\|^2 + \|\hat{\boldsymbol{z}} - \overline{\boldsymbol{z}}\|^2 + 2\langle\hat{\boldsymbol{x}} - \overline{\boldsymbol{x}}, \hat{\boldsymbol{z}} - \overline{\boldsymbol{z}}\rangle}{\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|^2}$$

$$= \frac{\boldsymbol{SSE_x}}{\boldsymbol{SST}} + \frac{\boldsymbol{SSE_z}}{\boldsymbol{SST}} + \frac{\boldsymbol{SSE_{xz}}}{\boldsymbol{SST}} = \boldsymbol{R_x^2} + \boldsymbol{R_z^2} + \boldsymbol{R_{xz}^2}$$

which I interpret as the proportions of total variation explained by the first group of predictors, the second, and their interactions, respectively.

4.6. References

Ables, G.P., Hens, J.R., and Nichenametla, S.N. (2016). Methionine restriction beyond life-span extension. Annals of the New York Academy of Sciences *1363*, 68-79.

Crider, K.S., Yang, T.P., Berry, R.J., and Bailey, L.B. (2012). Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. Advances in nutrition *3*, 21-38.

Kraus, D., Yang, Q., Kong, D., Banks, A.S., Zhang, L., Rodgers, J.T., Pirinen, E., Pulinilkunnil, T.C., Gong, F., Wang, Y.C., et al. (2014). Nicotinamide N-methyltransferase knockdown protects against diet-induced obesity. Nature *508*, 258-262.

Levine, M.E., Suarez, J.A., Brandhorst, S., Balasubramanian, P., Cheng, C.W., Madia, F., Fontana, L., Mirisola, M.G., Guevara-Aguirre, J., Wan, J., et al. (2014). Low protein intake is associated with a major reduction in IGF-1, cancer, and overall mortality in the 65 and younger but not older population. Cell metabolism *19*, 407-417.

Locasale, J.W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. Nature reviews. Cancer *13*, 572-583.

Mentch, S.J., Mehrmohamadi, M., Huang, L., Liu, X., Gupta, D., Mattocks, D., Gomez Padilla, P., Ables, G., Bamman, M.M., Thalacker-Mercer, A.E., et al. (2015). Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. Cell metabolism *22*, 861-873.

Shestov, A.A., Liu, X., Ser, Z., Cluntun, A.A., Hung, Y.P., Huang, L., Kim, D., Le, A., Yellen, G., Albeck, J.G., et al. (2014). Quantitative determinants of aerobic glycolysis identify flux through the enzyme GAPDH as a limiting step. eLife *3*.

Shiraki, N., Shiraki, Y., Tsuyama, T., Obata, F., Miura, M., Nagae, G., Aburatani, H., Kume, K., Endo, F., and Kume, S. (2014). Methionine metabolism regulates maintenance and differentiation of human pluripotent stem cells. Cell metabolism *19*, 780-794.

Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wagele, B., Altmaier, E., Deloukas, P., Erdmann, J., Grundberg, E., et al. (2011). Human metabolic individuality in biomedical and pharmaceutical research. Nature *477*, 54-60.

Thalacker-Mercer, A., Stec, M., Cui, X., Cross, J., Windham, S., and Bamman, M. (2013). Cluster analysis reveals differential transcript profiles associated with resistance training-induced human skeletal muscle hypertrophy. Physiological genomics *45*, 499-507.

Thalacker-Mercer, A.E., Petrella, J.K., and Bamman, M.M. (2009). Does habitual dietary intake influence myofiber hypertrophy in response to resistance training? A cluster analysis. Applied physiology, nutrition, and metabolism = Physiologie appliquee, nutrition et metabolisme *34*, 632-639.

Ulanovskaya, O.A., Zuhl, A.M., and Cravatt, B.F. (2013). NNMT promotes epigenetic remodeling in cancer by creating a metabolic methylation sink. Nature chemical biology *9*, 300-306.

CHAPTER 5: IDENTIFYING GENE EXPRESSION SIGNATURES OF RESPONSE TO ANTIMETABOLITE CHEMOTHERAPIES[8]

5.1 Abstract

Chemotherapeutic agents that target cellular metabolism are widely used in the clinic and are thought to exert their anti-cancer effects mainly through non-specific cytotoxic effects. However, patients vary dramatically with respect to survival and cancer outcome in response to treatment with these antimetabolite agents. Specific sources of this heterogeneity remain largely unknown. A deeper understanding of the extent that molecular information encoded in the metabolic pathways targeted by antimetabolite agent can explain variability in response is lacking. Here, I introduce a method for identifying gene expression signatures of response to chemotherapies and apply it to human tumors as well as cancer cell lines. This approach to analyzes genome-wide expression in an unbiased manner and identifies distinct favorable and unfavorable metabolic expression signatures. Importantly, metabolic pathways targeted by antimetabolites are enriched in the expression signatures. Finally, I characterize seventeen antimetabolite agents in various contexts and demonstrate that unlike common notion about their non-specific cytotoxicity, in fact specific metabolic factors explain variation in sensitivity to these agents.

[8] Manuscript under preparation for publication. Authors: Mehrmohamadi M, Jeong S.H., Locasale J.W.

5.2. Introduction

Cancer cells adapt their metabolism to meet the requirements of inappropriate growth, survival and proliferation (Pavlova and Thompson, 2016; Schulze and Harris, 2012). Since these demands are often not present in non-neoplastic cells to the same extent, there is considerable interest in exploiting these alterations for therapeutic advances. Antimetabolite chemotherapies are one of the most commonly used therapeutic strategies for the treatment of neoplastic disease (Chabner and Roberts, 2005). Historically, some of the first successful chemotherapeutic agents were derived from intermediates in the synthesis of folates (Farber, 1949; Farber and Diamond, 1948). Subsequently, there are now at least seventeen agents approved in the United States that target a specific metabolic enzyme (Cheung-Ong et al., 2013). These agents can often be tolerated and can achieve remarkable responses in advanced stage cancers leading to complete remission in many cases. However, the clinical responses to these agents are heterogeneous with some patients exhibiting resistance.

There is little molecular level information that is used clinically for prognostication. For instance, 5-fluorouracil (5-FU) is a widely used antimetabolite chemotherapy that interferes with pyrimidine biosynthesis by targeting the enzyme thymidylate synthetase (TYMS). Previous studies of association between cellular levels of TYMS and tumor response to 5-FU have been controversial, and currently TYMS expression is not used as a biomarker in clinical decision-making (Showalter et al., 2008). Other studies have found *TP53* mutational status as the only strong predictor of 5-FU outcome (Iorio et al., 2016; Kandioler et al., 2015). However, it

remains unclear whether the activity of the specific pathway that is targeted by 5-FU has any association with its impact on tumors. Other agents such as Methotrexate and Gemcitabine also have specific targets within one-carbon metabolism and nucleotide metabolism, but no successful biomarkers of response to these agents within the metabolic network are known.

The wealth of genomic information on annotated tumors now publically available through the cancer genome atlas (TCGA) allows these questions to be addressed in a more systematic way than previously possible. One study applied an unbiased analysis of genomic data on the TCGA ovarian cancer tumors and specifically looked for prognostic markers of response to Cisplatin using progression free survival of recipients, and was able to identify novel genetic and epigenetic subgroups with variable outcome (Hsu et al., 2012). Despite difficulties in studying drug response in human patients in the presence of numerous confounding factors and heterogeneity in therapeutic regimens, the unbiased framework introduced in that study provided useful insights (Hsu et al., 2012). This motivated us to apply a similar approach to identify gene expression subgroups of response to antimetabolite chemotherapies.

Unlike limitations associated with studying drug response in patients, cell line studies of drug response are more straightforward in some ways as cells are cultured in highly controlled laboratory environments, and dose-response analysis allows for careful quantification of sensitivity. A recent study used a large panel of cell lines from the catalog of somatic mutations in cancer (COSMIC) collection and

characterized molecular markers of response to hundreds of different drugs (Iorio et al., 2016). This drug panel included a number of antimetabolite chemotherapies including MTX, 5-FU, and Gemcitabine, together with a number of other agents grouped as "cytotoxic drugs". This study comprehensively evaluated thousands of molecular features in their ability to act as predictive markers of sensitivity and found the *TP53* mutational status as the most dominant marker for antimetabolite agents mentioned above. For 5-FU, a handful of copy number variants (CNVs) were also found to be predictive of cell line resistance (Iorio et al., 2016). However, this study did not explore gene expressions beyond expression of only 11 popular pathways, which found no significant predictors. It remains to be investigated whether any differences among antimetabolite agents can be captured in gene expression signatures of response to each, and whether such gene expression signatures can add to our power of distinguishing subtypes with heterogeneous outcome.

Here, I carry out an investigation of a set of antimetabolite compounds that target metabolic enzymes for use in cancer therapy. These agents target different pathways including folate synthesis, nucleotide metabolism, and glutathione biosynthesis. I first develop an unbiased approach to identify gene expression signatures of response in patients. Subsequently, I consider cell line analysis as a complementary approach to assess markers of cell line sensitivity to antimetabolite agents. Together, my results demonstrate specificity in molecular markers and identify metabolic determinants of response to these agents.

5.3. Results

5.3.1 Gene expression signatures of response to antimetabolite chemotherapy in patients are enriched for metabolic pathways

To identify gene expression signatures associated with response to chemotherapies in patients, I undertook an unbiased genome-wide approach based on step-wise filtering adapted from the framework previously introduced by Hsu et al. (Hsu et al., 2012) (Fig. 5.1A). I used the TCGA for the source of my clinically annotated human tumor genomic data. Progression free survival (PFS) was used as a measure of patient response to chemotherapy. TCGA cancer types in which patients were treated with a common antimetabolite agent were considered if both RNA-seq gene expression and follow-up data were available for a large enough cohort of patients (N>50). Since my goal was to identify subtypes of cancer patients with "good response" and "poor response", I considered each cancer type separately. These criteria limited my analyses of human data to 5-FU response in colon cancer and Gemcitabine response in pancreatic cancer.

5.3.1.1 Response to 5-FU in colon cancer

A total of 109 colon cancer patients were considered who received adjuvant 5-FU therapy as part of their chemotherapy combination. To avoid bias in my analysis of gene expression, I considered all of the genes in the genome after filtering out low-count mRNA expressions (Methods). I first calculated association between expression of each gene with PFS using univariate Cox regression (Methods), and filtered out

genes that did not show a significant ($p<0.05$) association. Next, I looked at the remaining 446 genes and further filtered out stage, age, *TP53* mutation, and nodal status associated genes to eliminate confounding factors that might affect association of genes with 5-FU response (see Methods). This lead to a final set of 299 genes that were each individually significantly associated with patient response to 5-FU in colon cancer, and their relationship to PFS was independent of stage, age, *TP53* mutation, and nodal status of the tumors (Figure 5.1A). Notably, this set included TYMS—the direct target of 5-FU. Next, I assessed the power of TYMS expression alone in distinguishing response subgroups. For this, I divided tumors into two groups based on their TYMS expression level: "low-TYMS" and "high-TYMS" (see Methods). I then compared PFS between the two groups using Cox regression and found a modestly significant difference in response between the low-TYMS and high-TYMS groups ($p=4.9e10\text{-}2$; Figure A3.S1A). Given that adjuvant 5-FU therapy is usually administered in stage III colon cancer, I repeated this analysis in stage III tumors only (N=59), and found a slightly stronger association ($p= 6e10\text{-}3$; Figure A3.S1B). In both analyses, I found that higher expression of TYMS is associated with poorer response to 5-FU therapy, consistent with previous reports (Hu et al., 2003; Wakasa et al., 2015), possibly explained by higher resistance of these tumors to TYMS inhibition.

I next set out to assess the combined power of all 299 genes in separating response subgroups. For this, I used a scheme previously proposed by Hsu et al. for DNA methylation (Hsu et al., 2012), and modified the method to apply to gene expression analysis (Methods). First, I converted the gene expression matrix into a discretized matrix of "favorability scores", where a gene with high expression in a

patient in the better prognosis subgroup was assigned a score of 1 (favorable), and a gene with high expression co-occurring with poorer prognosis subgroup was assigned a score of -1 (unfavorable), and all other cases were assigned a score of 0 (neutral) (see Methods for details). The clustering heatmap of the favorability scores discovered distinct subsets of genes (favorable vs. unfavorable) as well as distinct subgroups of patients (Figure 5.1B). To assess the functional relevance of the favorable and unfavorable gene signatures, I performed gene set enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. The unfavorable gene set was enriched for the following KEGG pathways: Circadian entrainment (p=7e10-3); Nucleotide sugar metabolism (p=7e10-3); Notch signaling (p=7e10-3); and One-carbon metabolism (p=1e10-2). TYMS, SHMT2, GALT, RENBP, and AMDHD2 were among the metabolic genes that had an unfavorable expression in colon cancer, meaning that their high expression in patients treated with 5-FU was associated with poorer prognosis. Consistent with my results, one-carbon metabolic fluxes have previously been shown to correlate with sensitivity to 5-FU in vitro and in mice (Ser et al., 2016). These observations illustrate the importance of specific metabolic target pathways of 5-FU in explaining part of the variability in patient response to this drug. Enrichment analysis on the favorable gene cluster showed enrichment of lipid metabolic KEGG pathways (Synthesis of unsaturated fatty acids (p=4e10-4); and Fatty acid metabolism (p=2e10-3)), with SCD and ACOX1 fatty acid de-saturases being among the metabolic genes in this group. Lipid synthesis has long been known to increase upon carcinogenesis, producing cellular membrane subunits for rapidly proliferating cells (Beloribi-Djefaflia et al., 2016). However, lipidome analyses have

shown that the role of fatty acids in cancers are more complex, with an enrichment of saturated fatty acids causing the loss of membrane fluidity, increase in drug resistance, and increase in malignancy of cancer cells (Rysman et al., 2010). My results confirm previous studies by identifying fatty acid oxidases and de-saturases SCD and ACOX1 as favorable enzymes, implicating a role for fatty acid metabolism.

To compare the two patient subgroups identified by this approach, I performed k-means clustering on the matrix of favorability scores and identified a distinct subgroup enriched with favorable genes (Group 1 in Figure 5.1B), and a second subgroup enriched with unfavorable gene expression (Group 2 in Figure 5.1B) (see Methods; Figure 5.1B). When PFS was compared between these two subgroups, I found a highly significant difference (Cox p= 3.46e-07; Figure 5.1C). This result is interesting as it shows that my scheme of discretizing combined gene expression signatures followed by favorability scoring and clustering is able to identify prognosis subgroups that are significantly more distinct than the subgroups identified based on TYMS expression alone, despite TYMS being the direct target of 5-FU. Importantly, my gene expression signatures are not associated with other prominent clinical predictors of prognosis (e.g. age, stage, nodal status, and *TP53* mutation), as I controlled for these confounding factors in the gene selection step (Methods; Figure 5.1A). This suggests that the gene expression signatures identified here offer additional information about prognosis beyond what is already captured by commonly used clinical metrics. Results also suggest that metabolism is an interconnected network of reactions that work in concert together; thus the combined activity of

multiple connected genes and pathways is a better reflection of the biological state of a tumor than the activity of individual enzymes.

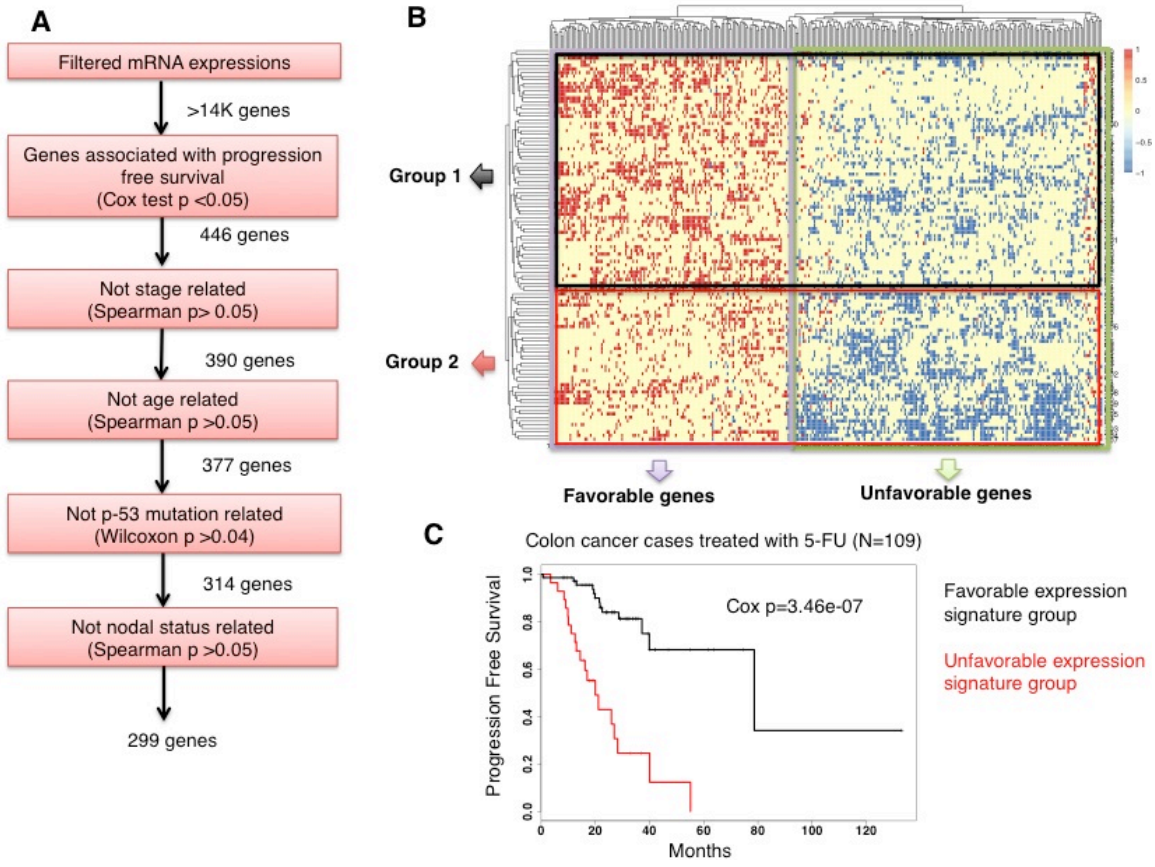


Figure 5.1— Combined gene expression signatures of response to 5-FU in colon cancer identify novel subgroups.

A) Schematic of the step-wise filtering used for gene selection in colon cancer (TCGA COAD).

B) Hierarchical clustering of heatmap of the discretized gene favorability scores. Columns represent genes and rows represent individuals. Favorable scores are shown by the color red (F=1), unfavorable by blue (F= -1), and neutral by yellow (F=0) (see Methods).

C) Kaplan-Meier plot showing progression free survival in the two tumor subgroups identified in part B.

5.3.1.2 Response to Gemcitabine in pancreatic cancer

I next set out to apply my gene expression analysis method to an independent TCGA cohort consisting of pancreatic cancer patients (N=100) who were treated with adjuvant Gemcitabine chemotherapy as part of their chemotherapy regimen. Gemcitabine is another chemotherapeutic agent that targets nucleotide and glutathione metabolism. My gene selection and filtering steps resulted in a set of 665 genes associated with PFS in this cohort after controlling for patient age, tumor grade, and *TP53* mutational status (Figure 5.2A). Visualization of a discretized expression heatmap made apparent subsets of favorable and unfavorable genes (Figure 5.2B). Pathway analysis of the favorable gene set showed Glycerophospholipid metabolism (p=1e10-4) pathway being enriched, while the following KEGG pathways were enriched in the unfavorable expression signature: Mitotic cell cycle and nuclear division (p<10e-9), Viral carcinogenesis (p=2e10-4), Mismatch repair (p=2e10-4), Apoptosis (p=8e10-3), and Pyrimidine metabolism (p=1e10-2) (Figure 5.2B). Notably, the unfavorable gene set included ribunucleotide reductases RRM1 and RRM2— direct targets of Gemcitabine— as well as DTYMK and TK1 in thymidine metabolism and NT5E in purine degradation pathways, demonstrating a role for specific target pathways of Gemcitabine in explaining response to this agent. The favorable gene signature included the following metabolic genes: *PLA2G2D, PLA2G4A, PLA2G4C,* and *PLD2* phospholipases, *LPGAT1, PNPLA6, AGPAT1,* and *AGPAT4*. This observation further supports previous cancer profiling studies that have established important structural and signaling roles for phospholipids in the pathogenesis and malignancy of cancer cells (Beloribi-Djefaflia et al., 2016).

I next performed k-means clustering on the matrix of favorability scores across these 665 genes and identified clear subgroups of patients. Comparison of the subgroup enriched with unfavorable gene expression with that of the favorable subgroup showed a significant difference in PFS (Cox p=1.8e-4; Figure 5.2C). Notably, when considered individually, RRM1 and RRM2 each had far less distinctive power (Cox p=6e10-3 for RRM1 and p=5e10-3 for RRM2; Figure A3.S2A,B) than the combined gene sets, further confirming the advantage of my approach by considering pathways rather than individual genes. Together, these results confirm the generalizability of this approach for identifying clinically distinct subgroups of cancer patients using gene expression signatures.

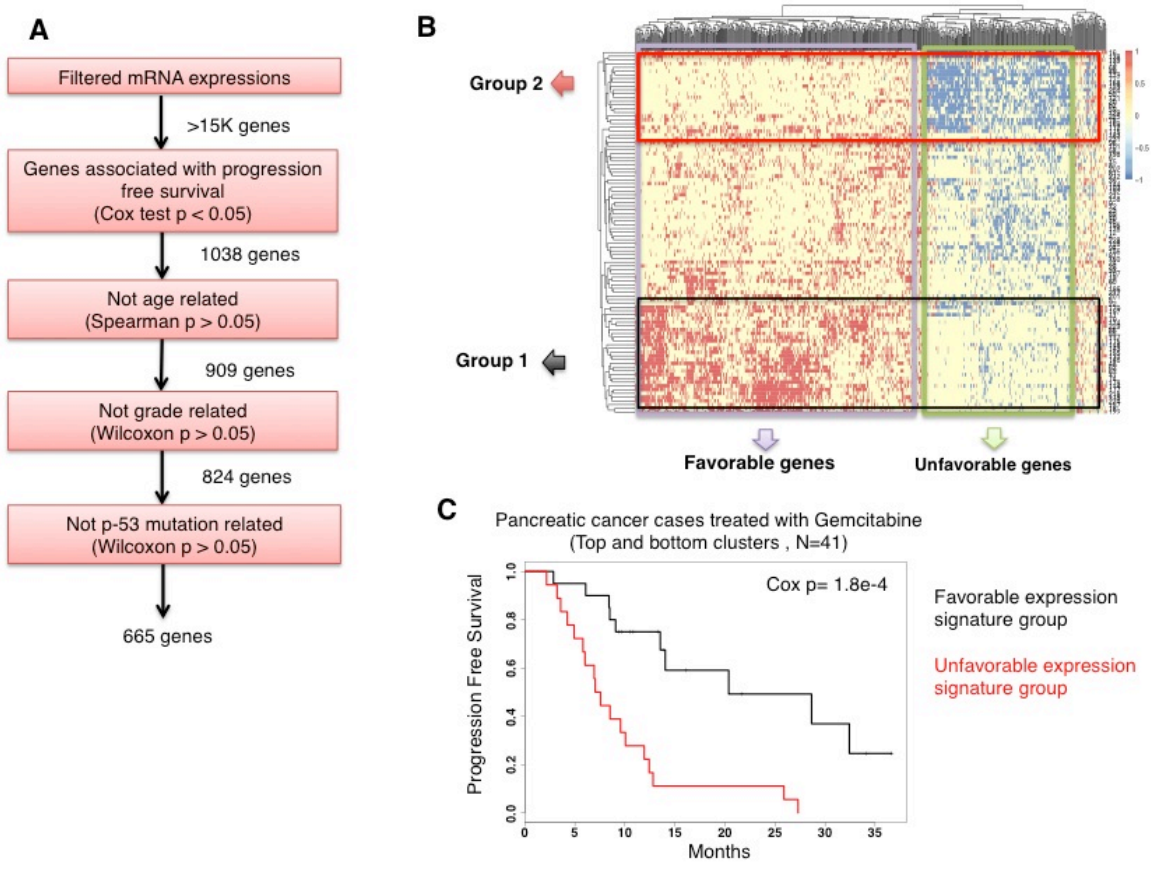


Figure 5.2— Combined gene expression signatures of response to Gemcitabine in pancreatic cancer identify novel subgroups.

A) Schematic of the step-wise filtering used for gene selection in pancreatic cancer (TCGA PAAD).

B) Hierarchical clustering of heatmap of the discretized gene favorability scores. Columns represent genes and rows represent individuals. Favorable scores are shown by the color red (F=1), unfavorable by blue (F= -1), and neutral by yellow (F=0) (see Methods).

C) Kaplan-Meier plot showing the progression free survival in the two tumor subgroups identified in part (B).

5.3.2 Specificity in gene expression signatures of cell line sensitivity to antimetabolite drugs

Due to limitations in the availability of sufficiently annotated human data with gene expression and follow-up information, I next turned to cancer cell lines to further test the applicability of my method. I used the catalog of somatic mutations in cancer (COSMIC) cell line set as the largest collection of annotated cancer cell lines and obtained microarray gene expression data as well as drug sensitivity information in the form of 50 percent of maximal inhibition of cell proliferation (IC-50) for the same agents I had previously tested in human samples (i.e. 5-FU and Gemcitabine). In the case of cell lines, I considered a gene favorable if its high expression co-occurred with higher sensitivity to drug treatment (lower IC-50), and unfavorable if its high expression co-occurred with lower sensitivity (higher IC-50) (see Methods).

5.3.2.1 Response to 5-FU in colon cancer cell lines

A set of 44 cell lines from colorectal origins was considered. For the gene selection step, I calculated correlation between expression of every gene in the genome with IC-50 value for 5-FU, and selected genes with a Kendall's tau value of 0.2 or larger and a corresponding p-value of 0.01 or smaller. A total of 364 genes passed this filter (Figure 5.3A). Subsequently, the discretization and favorability scoring approach as described in the previous section was applied to this matrix and the clustering heatmap was visualized (Figure 5.3B). Distinct subsets were immediately obvious, with favorable genes enriched in Protein processing (p=4e10-5), Arginine and proline metabolism (p=7e10-3), and glutathionie metabolism (p=8e10-3), while the unfavorable genes were not significantly enriched in any of the KEGG pathways. Dihydropyrimidine dehydrogenase (DPYD) was the only metabolic gene identified in the unfavorable set, consistent with its biological function and previous reports of its predictive power in 5-FU treated rectal cancers (Huang et al., 2013).

Next, I compared response to 5-FU between the two subgroups of cell lines identified by k-means clustering of the favorability matrix. The subgroup of cells enriched with the unfavorable gene expression signature had a significantly higher IC-50 for 5-FU (higher resistance), than the subgroup enriched with the favorable signature (Wilcox test p= 1.96e-11; Figure 5.3C). Together, these results confirm the generalizability of this method for identification of novel subgroups with distinct response to 5-FU, and also find a specific metabolic target (DPYD) as a marker of cell line sensitivity.

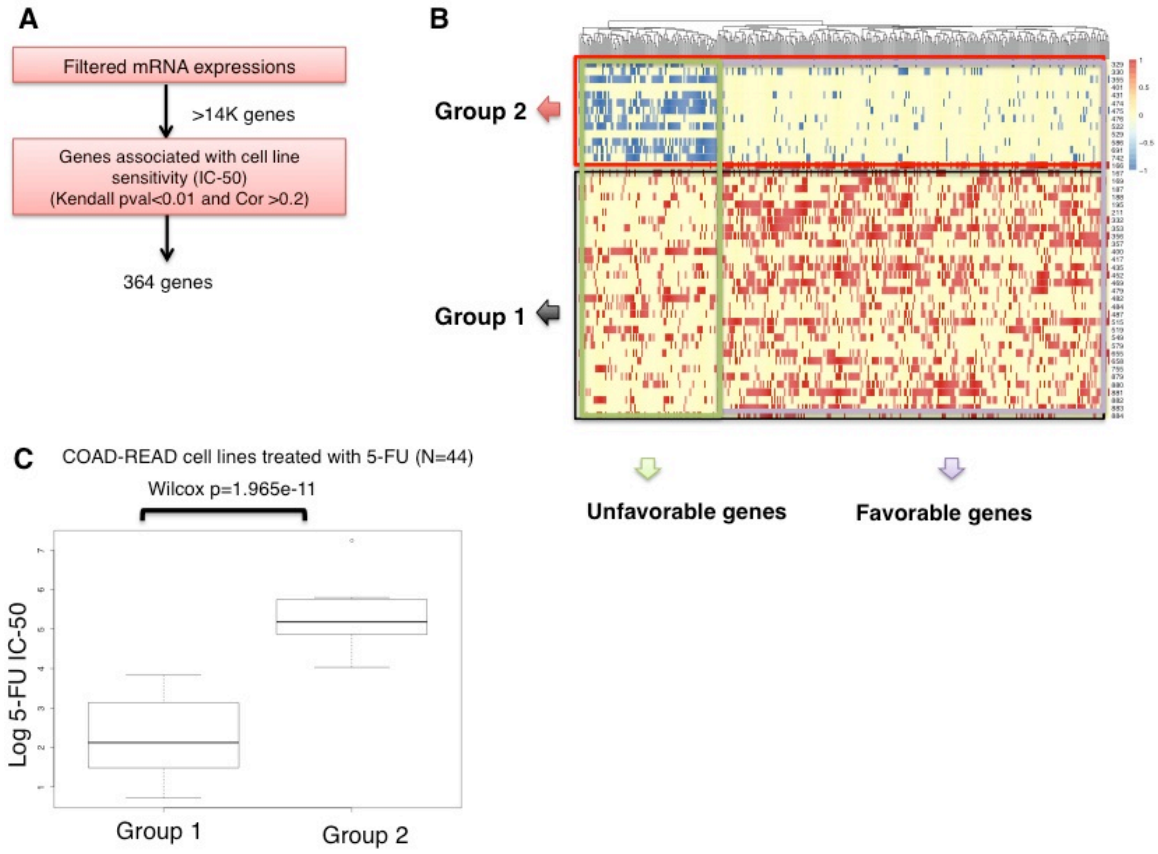


Figure 5.3— Combined gene expression signatures of response to 5-FU across colon cancer cell lines identify novel subgroups.

A) Schematic of the step-wise filtering used for gene selection in colon cancer (COSMIC COAD-READ).

B) Hierarchical clustering of heatmap of the discretized gene favorability scores. Columns represent genes and rows represent individuals. Favorable scores are shown by the color red (F=1), unfavorable by blue (F= -1), and neutral by yellow (F=0) (see Methods).

C) Box-plots comparing the resistance to 5-FU (log IC-50 values) between the two cell line subgroups identified in part (B) (error bars show the range of the data points in each group).

5.3.2.2 Response to Gemcitabine in pancreatic cancer cell lines

I next considered all COSMIC cell lines derived from pancreatic origins regarding their sensitivity to Gemcitabine. This set included only 17 cell lines, limiting the statistical power of this analysis. Only 201 genes passed my initial filtering (Figure A3.S3A). A visualization of the favorability heatmap illustrated two distinct clusters of genes, one with a mostly favorable expression score, but the second one with heterogeneous scores across the cell lines (Figure A3.S3B). Pathway analysis of the favorable set identified Chemical carcinogenesis (p=7e10-3), glutathionie metabolism (p=2e10-2), and Drug metabolism (p=4e10-2) KEGG pathways significantly enriched, while the unfavorable set was enriched in Adherens junctions (p=5e10-3), Bacterial invasion (p=6e10-3), and Glycophospholipid synthesis (p=7e10-3). Finally, comparison of sensitivity to Gemcitabine between two of the cell line subsets with distinct signatures found a significant difference in IC-50 (Wilcox p-value= 8e10-4; Figure A3.S3C), showing the power of this approach even when applied to such small data sets.

5.3.3 Analysis of metabolite profiles of cell lines identifies variability in metabolic markers

So far my results have shown considerable contribution from the metabolic gene expression network in distinguishing drug response subsets within human tumors as well as cancer cell lines. Careful consideration of two nucleotide metabolism inhibitors — 5-FU and Gemcitabine —revealed subtle differences in gene expression signatures associated with favorable and unfavorable response in each case, suggesting that even closely similar cytotoxic agents may act through different cellular pathways

and therefore be explained by different markers. My approach used gene expression levels of metabolic enzymes as surrogates for metabolic fluxes or enzyme activities. Next, to gain a more direct measure of metabolic states, Seong H. Jeong — an undergraduate student that I supervised— attempted to complement my analyses by taking advantage of direct metabolite consumption and release measurements across a panel of 60 cancer cell lines. The metabolic activities in the form of consumption or release rates (CORE) (Jain et al., 2012) were correlated with IC-50 values of 17 antimetabolite compounds (see Methods for the complete list; Figure 5.4A). Interestingly, the release rate of phosphocholine showed a strong negative correlation with sensitivity to 6 of the antimetabolite agents tested (Figure 5.4A). This result suggests that cells that have a higher rate of phosphocholine production are less sensitive to drug treatments, consistent with my gene expression results showing the enrichment of phospholipid metabolic genes in response signatures. Similarly, previous studies have shown that an increase in phosphatidylcholine affects cancer cell membrane dynamics and correlates with higher tumor malignancy and poorer overall survival (Beloribi-Djefaflia et al., 2016). Jeong's results agree with previous reports suggesting that increased amounts of phosphatidylcholine in the membrane of cancer cells may protect them from the uptake and toxicity of drugs, while high activity of enzymes that degrade phosphatidylcholine renders cells more sensitive to drug treatments, potentially contributing to a more favorable outcome for chemotherapy (Beloribi-Djefaflia et al., 2016). An example of a specific interaction that was detected at the level of metabolite consumption and release is sensitivity to Fludarabine— a purine analog— that was significantly associated with CORE of 2-deoxycytidine

(Figure 5.4A). Together, these results identify relationships between directly measured metabolic signatures of cancer cells and their sensitivity to antimetabolite chemotherapies, and also demonstrate variability among the 17 antimetabolites tested regarding their interaction with cellular metabolism.

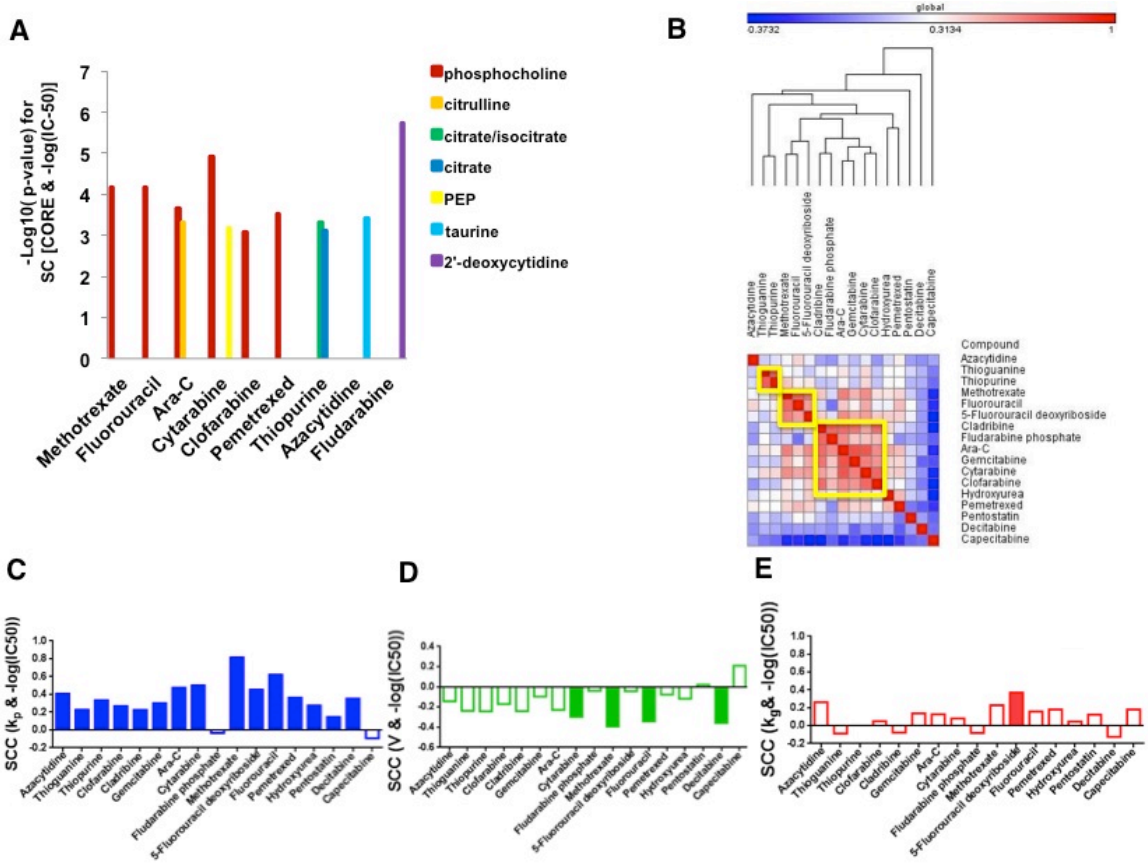


Figure 5.4— Analysis of additional determinants of sensitivity to antimetabolite agents demonstrate variability among these agents. [9]

A) The significance of association between metabolic profiles (consumption and release rates (CORE)) and sensitivity to drugs (-log (IC-50)) was assessed using spearman correlations (SC) across the NCI-60 cell line panel. The y-axis shows negative log-10 of the corresponding correlation p-values for only the significant associations found (q-value < 0.05).

B) Hierarchical clustering of the Pearson similarity matrix between the IC-50 values of 17 antimetabolite agents across the NCI-60 panel. The diagonal shows correlation of each drug with itself (=1). The yellow boxes show three distinct clusters of drugs.

[9] Credit goes to Seong H. Jeong.

C) Spearman correlation coefficient (SCC) between proliferation rate (kp) and sensitivity to each drug (-log (IC-50)) is shown. Solid bars show significant correlations (FDR-corrected q-value <0.05).

D) Spearman correlation coefficient (SCC) between cell volume (V) and sensitivity to each drug (-log (IC-50)) is shown. Solid bars show significant correlations (FDR-corrected q-value <0.05).

E) Spearman correlation coefficient (SCC) between growth rate (kg) and sensitivity to each drug (-log (IC-50)) is shown. Solid bars show significant correlations (FDR-corrected q-value <0.05).

5.3.4 Sensitivity to antimetabolite agents is more specific than previously appreciated

The gene expression results suggest that despite common cytotoxic effects of antimetabolite agents, they might have distinct biological markers in cells that are specific to the target metabolic pathways. Also, the analysis of metabolic CORE profiles in cell lines suggested that markers of sensitivity to antimetabolite agents might be more variable than previously appreciated. This motivated us to further assess specificity of determinants of response across a larger set of antimetabolite agents. We considered a set of 17 antimetabolite chemotherapeutic compounds (see Methods for the complete list). These agents target enzymes involved in a number of metabolic pathways including de novo nucleotide metabolism, amino acid metabolism, and glutathione metabolism. To assess the extent of correlation in the sensitivities of cell lines to these compounds, Jeong computed a similarity matrix of pairwise Pearson correlations between the IC-50 values for antimetabolite. Three distinct clusters were identified by hierarchical clustering: a cluster including Thiopurine and Thioguanine, a cluster for an anti-folate Methotrexate (MTX) and

pyrimidine analogs (5-FU and 5-FUDR), and a cluster for other purine analogs (Figure 5.4B). The antimetabolite compounds in the second cluster shared TYMS as a target enzyme. This analysis suggests that in general, compounds with common mechanisms of action tend to have similar sensitivity profiles across cell lines.

A common notion is that cytotoxicity of antimetabolite chemotherapies occurs in all rapidly dividing cells and thus lacks specificity. It has also been proposed that cell size, cell proliferation, and cellular metabolism are invariably coupled (Dolfi et al., 2013). Given that data on proliferation rate, cell size, and metabolic profiles are readily available for the NCI-60 cell lines, we sought to re-investigate these relationships in the context of association with cell line sensitivities to antimetabolite agents. Spearman rank correlations between IC-50 and proliferation rate were computed and revealed significant positive correlations (q-values< 0.05 in all cases except Capecitabine and Fluodarabine phosphase; Figure 5.4C). When the cell volumes were correlated with responses to antimetabolites, all compounds except for Capecitabine showed a negative correlation (four compounds had q-value<0.05) (Figure 5.4D). Together, these results confirm that cytotoxicity, as defined as the concentration of drug needed to achieve toxic dosages, is lower with smaller cells that also tend to divide more rapidly due to their size (Dolfi et al., 2013). The significant negative correlation between proliferation rate and cell volume suggested that to obtain an overall growth rate corresponding to the rate of synthesis of macromolecules, the proliferation rate should be corrected for cell volume (see Methods). Jeong next correlated dose responses with the volume-corrected proliferation rate, referred to hereinafter as the "growth rate" (Figure 5.4E). The strong

correlations that were observed between IC-50 values and proliferation rate were absent when considering the growth rates (Figure 5.4E). Thus, although cytotoxicity of antimetabolite agents appears highly non-specific with selectivity pertaining only to proliferation rate, these effects are completely removed when considering an overall growth rate. This suggests that unlike the common notion, variation in sensitivity to antimetabolite agents is not explained by differences in the actual rates of production of macromolecules in cells.

5.4 Discussion

The specificity of antimetabolite chemotherapeutic agents has long been questioned and previous reports have not reached a consensus about a prognostic value for expression levels of target enzymes for most of these agents. Given that the metabolic network is composed of complex interactions between multiple enzymes and pathways, we hypothesized that perhaps by defining gene signatures instead of individual enzyme markers, we would gain power in distinguishing subgroups of tumors with differential response to therapy.

Here, I introduced an unbiased approach for the assessment of combined prognostic power of expression of multiple genes and used this platform to define favorable and unfavorable signatures. Notably, I showed that these signatures allow for distinguishing novel poor prognosis (high progression rate) from good prognosis (low progression rate) subgroups far more robustly than individual target genes.

Importantly, since the gene selection steps control for expression differences related to other important clinical and genetic attributes of response, I am assured that the gene signature analysis captures information about response subgroups beyond the already established markers.

In both studied cases of 5-FU in colon cancer and Gemcitabine in pancreatic cancer, I found that expression of metabolic pathways related to direct targets of the drugs are enriched in the unfavorable gene set. This confirmed that tumors with higher activity of target pathways require higher doses of drug to elicit the inhibitory response and are therefore more resistant to treatment. However, my results discovered that metabolic state of cells are not fully reflected in the expression levels of individual target enzymes, but rather captured more robustly in the collection of functionally and chemically linked enzymes in pathways. Although I was only able to illustrate the applicability of my method in two independent cohorts of human tumors due to data limitations, results suggest generalizability of this method to other antimetabolite agents as well.

Gene signatures associated with favorable and unfavorable response for 5-FU and Gemcitabine exhibited functional similarities overall, but distinct markers for each drug were also discovered. In both cases of 5-FU and Gemcitabine, high expressions of the target metabolic pathways (i.e. nucleotide metabolism) were associated with unfavorable outcome, while high expression of lipid metabolizing pathways were associated with favorable outcome. These point to common general mechanisms of cellular response to these drugs. However, a deeper look into specific genes and

pathway within the signatures for 5-FU and Gemcitabine identified some differences. For instance, while One-carbon metabolism and Nucleotide sugar metabolism were identified as the unfavorable signature for 5-FU, Pyrimidine metabolism was discovered in the case of Gemcitabine. Furthermore, *TYMS* was among the unfavorable genes for 5-FU, while *RRM1* and *RRM2* were among the unfavorable genes for Gemcitabine. Together, these results suggest that despite similarities in overall mechanisms of action, antimetabolite agents have specific biological markers that have not been very well characterized and appreciated in the past.

My parallel analyses of cancer cell line sensitivities to the same chemotherapeutic agents also proved useful in identifying distinct subgroups using the gene signature approach. Other than lipid metabolic genes, the gene sets identified as favorable and unfavorable signatures in cell lines did not completely match that identified from the analysis of PFS in human tumors. The main sensitivity predictor in vitro seemed to be Glutathione metabolism and Drug metabolism that were found in both cases of 5-FU and Gemcitabine to be associated with favorable outcome (i.e. higher sensitivity of cells to drug treatment). This observation is consistent with previous reports showing a critical role for glutathione metabolism in detoxification and protection against drugs in vitro (Traverso et al., 2013). These results illustrated that despite the availability and convenience of using cell lines as models of human tumors for drug response studies, analysis of patient tumors is advantageous in that it provides insights that are not fully reflected in cancer cell lines, potentially due to unwanted effects of culture media.

Together, our analyses of human tumors and cancer cell lines elucidated considerable variability among different antimetabolite agents, as well as specificity in metabolic markers of sensitivity to them. These demonstrate that despite the common notion, different classes of antimetabolite agents vary according to their distinct cellular functions. Our results suggest that potentially important biological markers of response to antimetabolite compounds exist, and a better understanding of these factors will provide useful insights for clinical decision-making. Notably, we showed that gene expression signatures have significant power to capture part of the previously unexplained variation in patients' responses to 5-FU and Gemcitabine in colon and pancreatic cancers, respectively. Future studies using larger cohorts of human tumors with well-annotated patient follow-up information can provide valuable additional insights about antimetabolite response signatures.

5.5 Methods

5.5.1 Survival analyses

5.5.1.1 Individual genes

TCGA progression free survival (PFS) across COAD and PAAD were obtained through the cBioPortal for cancer genomics. Level-3 RNA-seq RSEM gene-normalized counts were also downloaded through the GDC portal (https://gdc.cancer.gov/). The values were log2 normalized and in each data set, genes with a count of 2 or smaller in over 80% of the samples were removed as low-count genes. I used cancer progression as the "event" in Cox models and last day of follow-up to right censor the data in cases where no progression was documented. R packages "survival" was used for univariate survival analyses independently for all genes (Figure 5.1A and Figure 5.2A).

5.5.1.2 Gene signatures

When considering survival analysis for subgroups identified by my favorability scoring method (described in the following), I used the subgroup assignments based on the k-means clustering of favorability matrix in each case to label samples as "favorable signature group" and "unfavorable signature group". Subsequently, Cox regression was performed to assess the significance of the difference between PFS of the two groups as shown in Figure 5.1C and Figure 5.2C.

5.5.2 Cell line sensitivity analyses

For the COSMIC cell lines, RMA-normalized gene expressions were obtained through the Sanger Institute (http://cancer.sanger.ac.uk/cosmic). Genes with a coefficient of variation of 0.05 or smaller were removed. To test association with drug response, inhibitory concentration (IC-50) values were correlated with gene expression values and a Kendal tau was calculated. Genes with a correlation of over 0.2 and an associated p-value of 0.01 or less were selected for subsequent discretization step (Figure 5.3A and Figure A3.S3A).

5.5.3 Gene selection approach

Genes that passed my first filter i.e. showed a significant association with PFS (Cox p-value<0.05), were subsequently evaluated by additional clinical and genetic attributes. To eliminate genes whose expression levels were significantly affected by *TP53* mutational status, I compared expression levels in *TP53* mutant with *TP53* wild-type samples and a Wilcox non-parametric test was used to assess statistical difference. This test allowed filtering out genes significantly associated with *TP53* mutation. For other clinical attributes such as cancer stage, patient age, tumor grade, and nodal status, the Spearman correlation was used to test associations between gene expression and these clinical factors across samples. Finally, genes that passed all of the above filters were used for subsequent discretization analyses.

5.5.4 Survival analysis using expression of target enzymes

To assess the strength of direct target enzymes of 5-FU and Gemcitabine as markers of PFS, I considered expression levels of TYMS and RRM1 (RRM2), respectively. I first used the function "cutp" in the R package "survMisc" to find the best cutting point in the continuous gene expression. I the used this cutting point as a threshold to divide the samples into two groups of "low" and "high" expression for samples below and above the cut-point, respectively.

5.5.5 Discretizing gene expressions and defining favorability scores

I used the following criteria to discretize the signature gene expression matrix and label expressions "favorable" or "unfavorable" based on their relationship with PFS. A gene was assigned a value of 1 and was considered favorable if its high expression (higher than median plus half of the standard deviation for that gene) co-occurred with better prognosis, and a value of -1 (unfavorable) if its high expression co-occurred with poor prognosis in univariate Cox regression:

$$F \begin{cases} = 1, \text{ if } Eij \geq med+s/2 \quad \textbf{and} \quad j \in \text{good survival} \\ = -1, \text{ if } Eij \geq med+s/2 \quad \textbf{and} \quad j \in \text{poor survival} \\ = 0, \text{ otherwise} \end{cases}$$

, where Eij represents expression of gene "i" in individual tumor "j".

For discretizing cell line expression data, the following modified scheme was used where cell lines were labeled either "sensitive" or "resistant" to a drug if their IC-

50 value was at either extremes of the distribution of IC-50 values for that given drug across all cell lines.

$$F \begin{cases} = 1, & \text{if } Eij \geq med+s/2 \quad \textbf{and} \quad j \in \text{sensitive} \\ = -1, & \text{if } Eij \geq med+s/2 \quad \textbf{and} \quad j \in \text{resistant} \\ = 0, & \text{otherwise} \end{cases}$$

, where Eij represents expression of gene "i" in cell line "j".

5.5.6 Pathway enrichment analyses

Pathway enrichment analysis was performed on the resulting gene list for each cancer type using Enrichr (Chen et al., 2013). P-values from the Fisher's exact test are reported for significant ($p<0.05$) KEGG pathways.

5.5.7 Analyses of non-gene expression cell attributes

Jeong obtained IC-50 values for the 17 antimetabolite compounds across a panel of 60 cell lines from the National Cancer Institute (NCI-60) (Dolfi et al., 2013). To complement my gene expression analyses, we took advantage of the NCI-60 cell line panel where in addition to the comprehensive annotation of cell lines, a previous study has quantified the consumption and release rates (CORE) of hundreds of metabolite by each of these cell lines. Jeong obtained cell volumes, proliferation rates, CORE values, and dose response sensitivity information (IC-50 values) for 17 antimetabolite drugs across this cell line panel

(https://dtp.cancer.gov/discovery_development/nci-60/). CORE values are positive if a metabolite is released into the media by cancer cells and is negative if the metabolite is consumed. The list of these antimetabolic agents is as follows: Gemcitabine, Methotrexate, Pemetrexed, Thioguanine, Thiopurine, Fluorouracil, 5-Fluorouracil deoxyriboside, Hydroxyurea, Ara-C, Azacytidine, Cladribine, Decitabine, Pentostatin, Cytarabine, Fluodarabine phosphate, Clofarabine, and Capecitabine.

5.5.8 Growth rate calculations

We obtained growth rate by correcting proliferation rates for volumes. At time zero - right after the cell division, the cell volume (V_0) is the minimum. At time T_1, the cell gets bigger to V_1. If we define growth rate (k_g) as the increase of cell volume per time it takes, we can come up with the equation below:

$$\square = \boldsymbol{V_0 + T_1 k_g}$$

At doubling time (T_d), the cell will divide into two and we assume two divided cells will have the same volume as the initial volume, V_0.

$$\boldsymbol{V_2 = 2V_0 = V_0 + T_d k_g}$$

$$V_0 = T_d k_g \qquad -(1)$$

$$T_d = \frac{\ln 2}{k_p} \qquad -(2)$$

$$V_0 = \left(\frac{\ln 2}{k_p}\right) k_g - (1), (2)$$

Jeong then solved the above equation to obtain the following equation for growth rate:

$$k_g = \frac{V_0 k_p}{\ln 2}$$

5.6. References

Beloribi-Djefaflia, S., Vasseur, S., and Guillaumond, F. (2016). Lipid metabolic reprogramming in cancer cells. Oncogenesis *5*, e189.

Chabner, B.A., and Roberts, T.G., Jr. (2005). Timeline: Chemotherapy and the war on cancer. Nature reviews. Cancer *5*, 65-72.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics *14*, 128.

Cheung-Ong, K., Giaever, G., and Nislow, C. (2013). DNA-damaging agents in cancer chemotherapy: serendipity and chemical biology. Chemistry & biology *20*, 648-659.

Dolfi, S.C., Chan, L.L., Qiu, J., Tedeschi, P.M., Bertino, J.R., Hirshfield, K.M., Oltvai, Z.N., and Vazquez, A. (2013). The metabolic demands of cancer cells are coupled to their size and protein synthesis rates. Cancer & metabolism *1*, 20.

Farber, S. (1949). Some observations on the effect of folic acid antagonists on acute leukemia and other forms of incurable cancer. Blood *4*, 160-167.

Farber, S., and Diamond, L.K. (1948). Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid. The New England journal of medicine *238*, 787-793.

Hsu, F.H., Serpedin, E., Hsiao, T.H., Bishop, A.J., Dougherty, E.R., and Chen, Y. (2012). Reducing confounding and suppression effects in TCGA data: an integrated analysis of chemotherapy response in ovarian cancer. BMC genomics *13 Suppl 6*, S13.

Hu, Y.C., Komorowski, R.A., Graewin, S., Hostetter, G., Kallioniemi, O.P., Pitt, H.A., and Ahrendt, S.A. (2003). Thymidylate synthase expression predicts the response to 5-fluorouracil-based adjuvant therapy in pancreatic cancer. Clinical cancer research : an official journal of the American Association for Cancer Research *9*, 4165-4171.

Huang, M.Y., Wu, C.H., Huang, C.M., Chung, F.Y., Huang, C.W., Tsai, H.L., Chen, C.F., Lin, S.R., and Wang, J.Y. (2013). DPYD, TYMS, TYMP, TK1, and TK2 genetic expressions as response markers in locally advanced rectal cancer patients treated with fluoropyrimidine-based chemoradiotherapy. BioMed research international *2013*, 931028.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Goncalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740-754.

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., Kafri, R., Kirschner, M.W., Clish, C.B., and Mootha, V.K. (2012). Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. Science *336*, 1040-1044.

Kandioler, D., Mittlbock, M., Kappel, S., Puhalla, H., Herbst, F., Langner, C., Wolf, B., Tschmelitsch, J., Schippinger, W., Steger, G., et al. (2015). TP53 Mutational Status and Prediction of Benefit from Adjuvant 5-Fluorouracil in Stage III Colon Cancer Patients. EBioMedicine *2*, 825-830.

Pavlova, N.N., and Thompson, C.B. (2016). The Emerging Hallmarks of Cancer Metabolism. Cell metabolism *23*, 27-47.

Rysman, E., Brusselmans, K., Scheys, K., Timmermans, L., Derua, R., Munck, S., Van Veldhoven, P.P., Waltregny, D., Daniels, V.W., Machiels, J., et al. (2010). De novo lipogenesis protects cancer cells from free radicals and chemotherapeutics by promoting membrane lipid saturation. Cancer research *70*, 8117-8126.

Schulze, A., and Harris, A.L. (2012). How cancer metabolism is tuned for proliferation and vulnerable to disruption. Nature *491*, 364-373.

Ser, Z., Gao, X., Johnson, C., Mehrmohamadi, M., Liu, X., Li, S., and Locasale, J.W. (2016). Targeting One Carbon Metabolism with an Antimetabolite Disrupts Pyrimidine Homeostasis and Induces Nucleotide Overflow. Cell reports *15*, 2367-2376.
Showalter, S.L., Showalter, T.N., Witkiewicz, A., Havens, R., Kennedy, E.P., Hucl, T., Kern, S.E., Yeo, C.J., and Brody, J.R. (2008). Evaluating the drug-target relationship between thymidylate synthase expression and tumor response to 5-fluorouracil. Is it time to move forward? Cancer biology & therapy *7*, 986-994.
Traverso, N., Ricciarelli, R., Nitti, M., Marengo, B., Furfaro, A.L., Pronzato, M.A., Marinari, U.M., and Domenicotti, C. (2013). Role of glutathione in cancer progression and chemoresistance. Oxidative medicine and cellular longevity *2013*, 972913.
Wakasa, K., Kawabata, R., Nakao, S., Hattori, H., Taguchi, K., Uchida, J., Yamanaka, T., Maehara, Y., Fukushima, M., and Oda, S. (2015). Dynamic modulation of thymidylate synthase gene expression and fluorouracil sensitivity in human colorectal cancer cells. PloS one *10*, e0123076.

CHAPTER 6: CONCLUSIONS

6.1. Cancer as a heterogeneous disease

Due to instability and substantial heterogeneity, cancer is not considered one disease but a collection of multiple diseases. Variation exists between and within cancer types, between cancer subtypes, and even among clonal populations of cells within an individual tumor (McGranahan and Swanton, 2015). Such multi-level heterogeneity makes treatment of cancer exceptionally complex and challenging.

Though promising, personalization of therapy strategies for cancer patients requires a deeper understanding of various aspects of the heterogeneity in cancer (Chin et al., 2011). Altered metabolism is one of the relatively newly recognized features of cancer that is not yet fully characterized across different human cancers (Pavlova and Thompson, 2016). In this dissertation, I focused on the study of a specific metabolic network —one-carbon (1-C) metabolism— and provided insights on the heterogeneity with respects to various functions of 1-C metabolism in human cancers.

6.2. Summary of results

The main goal of my thesis was to characterize some of the important roles of one-carbon metabolism across human tumors in an unprecedented systematic

and quantitative manner. Findings from this work have significance in providing molecular insights about cancer metabolism, novel computational tools for analysis of high dimensional multi-layer genomic and epigenomic tumor data, and clinical and translational information toward personalized medication and dietary intervention in cancer patients.

In chapter 2, I studied the amino acid serine and its utilization across hundreds of human tumors. First, using a computational framework that I introduced, I estimated pathway activities from gene expression profiles of tumors. Comparative analyses at this level revealed significant heterogeneity among cancer types with respect to how they metabolize serine. Furthermore, co-regulation analysis revealed significant correlation between nucleotide and glutathione synthesis in all cancer types in the study, suggesting a common feature of cancers. I next used a panel of cancer cell lines to directly test the computational flux predictions. I used serine tracing in vitro followed by metabolomics. The experimental flux calculations agreed with my computational predictions in illustrating that the bulk of the serine flux is shunted toward nucleotide and glutathione synthesis in colon cancer. Furthermore, this study introduced a novel approach for estimating metabolic fluxes from pathway-level gene expression profile of tumors. This is significant as measuring metabolic fluxes in human tumors experimentally is not simple with current technology, especially for larger sample sizes. I showed that although expression levels of individual enzymes do not accurately reflect actual enzyme activities, combined gene expression information across metabolic pathways provide a reasonable estimate of pathway activities in tumors. Another important feature of flux distribution through one-carbon metabolism

that was identified by my computations and in vitro analyses was the de-coupled folate and methionine cycle fluxes. Together, my results revealed common as well as cancer-specific utilization of serine through one-carbon metabolism in a quantitative manner.

In the 3rd chapter, my main goal was to investigate the biochemical link between one-carbon metabolism and methylation by quantifying the extent that cancer DNA methylation profiles are predictable by the activity of one-carbon metabolism. To this end, I obtained multiplatform information on thousands of human tumors through the cancer genome atlas (TCGA) and applied computational modeling and machine-learning algorithms to quantify determinants of variation in DNA methylation between individual tumors of the same cancer type (inter-individual variation). I was able to show that a number of one-carbon metabolic enzymes— especially methionine adenosyl transferase (MAT) in the methionine cycle— have a significantly large predictive power of DNA methylation at multiple levels studied. Furthermore, my results illustrated that this relationship between one-carbon metabolism and DNA methylation exists at functionally important chromatin regions and overlaps with important cancer genes. Finally, I performed survival analyses and provided evidence for the clinical relevance of metabolic regulation of the DNA methylome by showing that loss of regulation of DNA methylation by the methionine cycle is associated with poorer overall survival. This is consistent with the model of cancer as a dysregulated epigenome that suggests loss of epigenetic stability provides cancer cells with higher plasticity and adaptive advantage (Timp and Feinberg, 2013).

In chapter 4, I assessed the physiological relevance of potentials for dietary intervention with the amino acid methionine in humans. Using dietary records as well as serum metabolomics on a cohort of human subjects, I studied methionine levels in the serum. Results confirmed that the physiological range of methionine encompasses the methionine concentrations previously shown to affect epigenetics. Furthermore, I used computational modeling to identify the determinants of serum methionine. My models showed a ~30% contribution from diet in explaining overall variation in serum methionine, ~30% for clinical factors (e.g. age, gender, etc), and ~20% unexplained, probably due to genetic differences between individuals. I also further identified the main food categories of dietary determinants of methionine in the serum. Together, these results confirm the relevance of potential epigenetic modifications through one-carbon metabolism using dietary intervention with methionine.

Finally in chapter 5, I focused on translating my findings about the important roles of 1-C metabolism in cancer into more clinically relevant discoveries by investigating patient response to antimetabolite chemotherapies. I introduced a novel approach for identifying gene expression markers of prognosis in an unbiased manner and showed that this approach can successfully separate novel poor prognosis from good prognosis subgroups of patients beyond the power of features already used in the clinic (e.g. tumor grade, stage, *TP53* mutation, etc). Furthermore, I illustrated that although the expression levels of individual enzyme targets of antimetabolite agents do not strongly associate with prognosis subgroups, the combined activity of multiple enzymes within a metabolic pathway is a strong predictor of therapy

outcome. Furthermore, using multiple independent datasets on primary human tumors as well as cancer cell lines, I showed that despite the common notion about the non-specific cytotoxic nature of most antimetabolite chemotherapies, there is considerable variability among different classes of such agents, and response to these agents strongly correlates with specific metabolic features of tumors suggesting potentials in utilizing these metabolic signatures as biomarkers for clinical decision-making in cancer therapy.

6.3. Limitations and future directions

My work provides a framework for the study of inter- and intra- cancer type metabolic heterogeneity using genomic profiles of tumors. I have analyzed over 8 different human cancer types and thousands of individual tumors within each type for the work presented in this thesis. In the future, larger datasets including additional cancer types and subtypes can be analyzed using the tools developed and described here, to fully represent the complete collection of human cancers and complement our current understanding of this field.

Furthermore, it is known that individual tumors are composed of multiple clonal populations that typically exhibit great variability. This is an important aspect of what makes cancer such a complex condition to understand and treat. The work presented here was performed using data on bulk tumor samples from individual patients, and therefore did not have sufficient resolution for analyzing within-tumor heterogeneity. In the future, by taking advantage of the analyses of multiple section

sub-sections from the same tumor or the novel technology of single cell genomics (Gawad et al., 2016), similar analyses can be performed with higher resolution using information about sub-clonal and cell-to-cell variability within a tumor. Understanding this level of heterogeneity is crucial for personalized medicine and informed decision-making, as well as understanding tumor evolutionary dynamics.

The bulk of the work presented here involved computational analyses and modeling of cancer using high dimensional multi-platform data on human tumors. In most cases, computational validation of models was reached using independent test sets. In the case of the serine fate study presented in chapter 2, the computational results were also complemented by experimental validation in cancer cell lines. Though rigorous computational cross-validation was used to assure the accuracy of in silico findings throughout the thesis, future experimental procedures that would directly test some of the main results presented here would provide further support for the role of 1-C metabolism in human cancers. Future studies on mechanistic details of the relationship between epigenetics and one-carbon metabolism, as well as potentials for dietary intervention with methionine in humans are needed to directly and mechanistically test some of my predictions.

Finally, the clinical translation of my findings in the area of identification of patient subgroups that would benefit from antimetabolite chemotherapies is a critical aspect. Current practice in the clinic does not involve use of metabolic markers for making decisions about prescribing antimetabolite

chemotherapies. I have provided evidence for potential advancements that can be made in this area. However, an important limitation in modeling drug response using the TCGA data is the lack of uniformity in therapy regimens as most patients received combination of multiple treatments. Furthermore, given that all TCGA samples had undergone adjuvant chemotherapy, progression in this group remains an indirect measure of response of the tumor to drug treatment with the assumption that remaining tumor cells after surgery represent the original tumor in patients. Although I was able to obtain interesting results in spite of these limitations and confounding factors inherent in the data, the field would benefit greatly from the availability of larger datasets with neo-adjuvant as well as adjuvant therapy cohorts and a more uniformly controlled chemotherapy combination since the unwanted variability in the current TCGA data substantially compromises the analysis and discovery power.

Despite limitations in the availability and types of data on human tumors, my thesis provides valuable computational framework and tools for studying complex human diseases using genomic information, sheds light on the biology of one-carbon metabolism in cancer, characterizes how different human cancers utilize this metabolic network toward their needs, quantifies the contribution from this metabolic network in cellular epigenetics, elucidates potentials for dietary intervention therapies using the intermediates in one-carbon metabolism, and demonstrates how these findings can be translated into clinically useful information.

6.4. References

Chin, L., Andersen, J.N., and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. Nature medicine *17*, 297-303.
Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nature reviews. Genetics *17*, 175-188.
McGranahan, N., and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer cell *27*, 15-26.
Pavlova, N.N., and Thompson, C.B. (2016). The Emerging Hallmarks of Cancer Metabolism. Cell metabolism *23*, 27-47.
Timp, W., and Feinberg, A.P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nature reviews. Cancer *13*, 497-510.

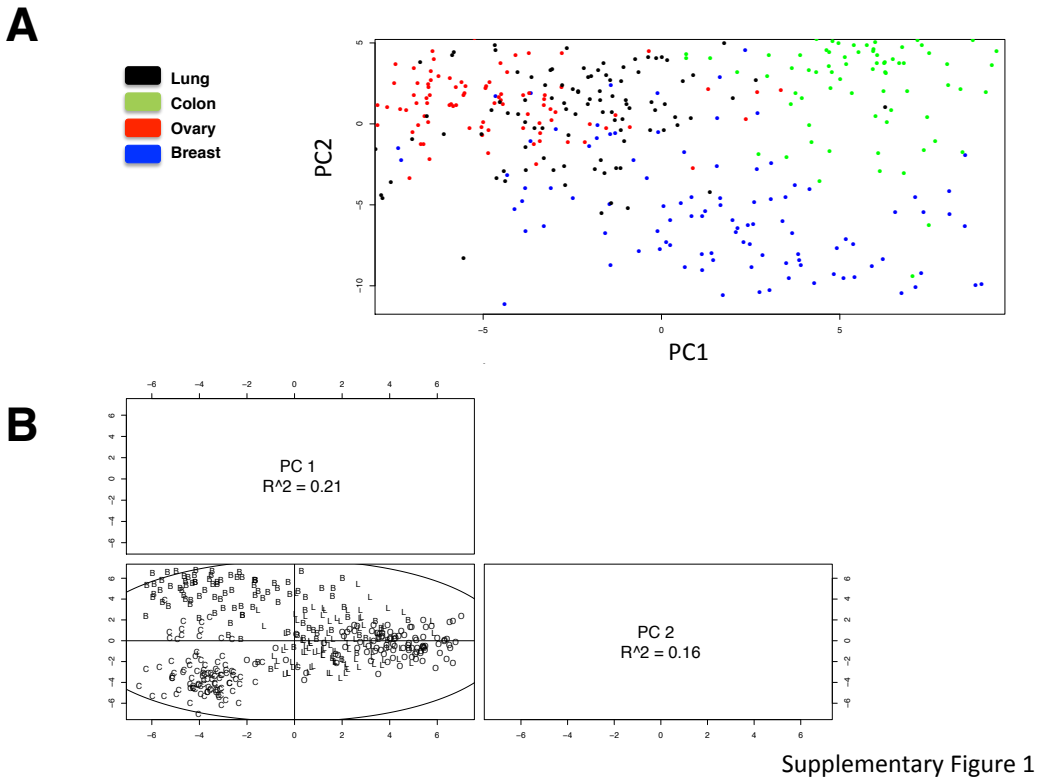
APPENDIX 1: SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Supplementary Methods.

Contingency table comparing mitochondrial vs. cytosolic isoforms in the context of tumor/normal over-expression (combined across all cancer types in study). This test was done only on enzymes that have both mitochondrial and cytosolic isoforms included in the SGOC network (6 mitochondrial and 5 cytosolic enzymes, 4 cancer types). Fisher's exact p-value = 0.02

	Over-expressed in tumor	Not over-expressed in tumor
Mitochondrial isoforms	12	12
Cytosolic isoforms	3	17

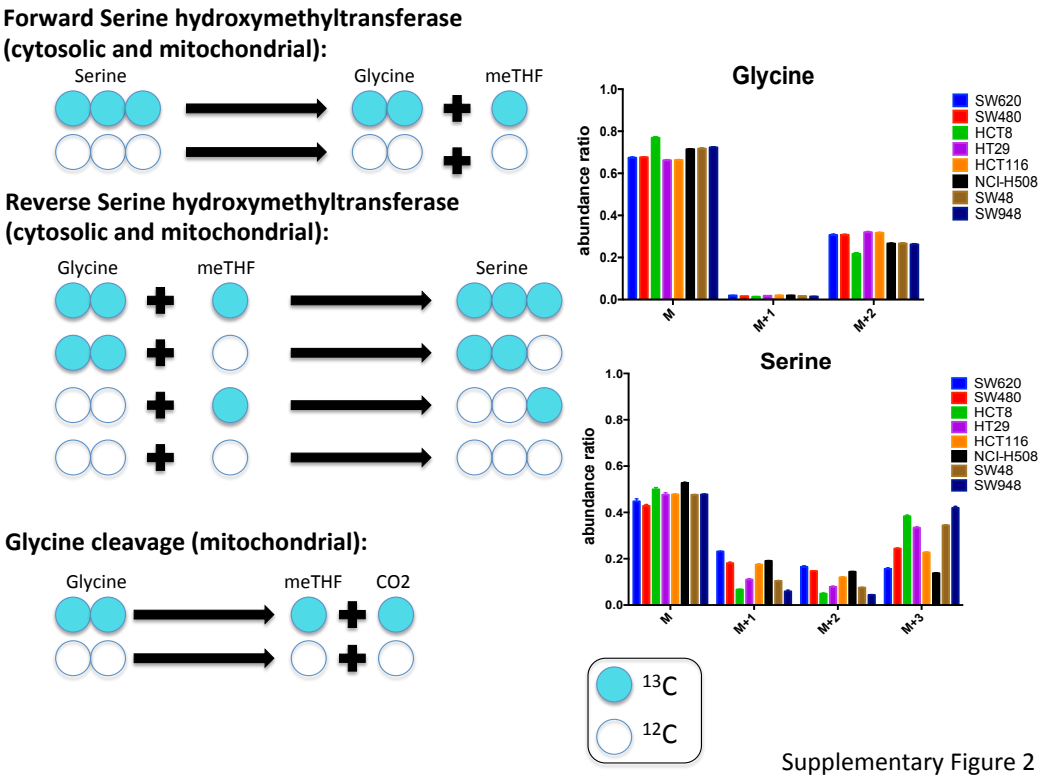
Figure A1.S1— Clustering of individual tumors based on gene expression profiles.



Supplementary Figure 1

A) Principal component analysis on 100 randomly picked samples from each cancer type (400 samples total). The second principal component (PC2) is plotted against the first principal component (PC1). B) Principal component analysis on 100 randomly picked samples from each cancer type (400 samples total). Different cancer types are shown by letters (B:Breast, O:Ovary, C:Colon, L:Lung). The second principal component (PC2) is plotted against the first principal component (PC1). Fraction of variation explained by each PC is shown as R^2.

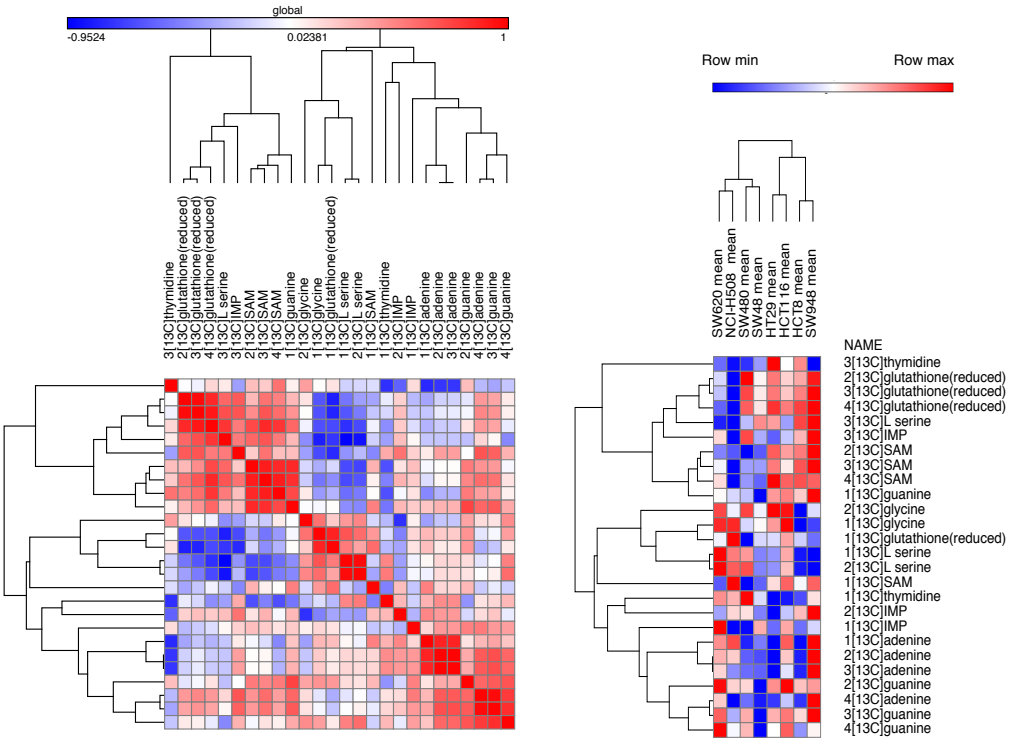
Figure A1.S2— Steady state labeling distribution on serine and glycine.



Supplementary Figure 2

Schematic showing the mass isotopomer distribution (MID) patterns observed for serine and glycine in my 13C-serine tracing experiment (right) along with the chemical reactions explaining the patterns observed (left).

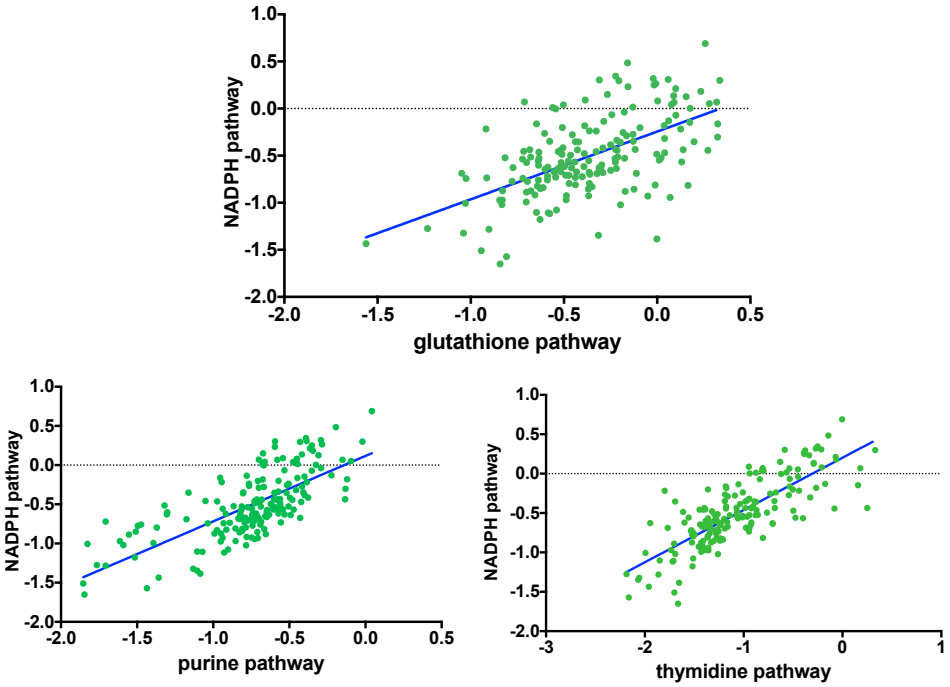
Figure A1.S3— Clustering of metabolite MIDs.



Supplementary Figure 3

Heatmaps of abundance ratios of some of the labeled isotopomers from the 13C experiment. On the right, hierarchical clustering is shown for the 8 colon cancer cell lines in the columns and the labeled metabolites in rows. The left heatmap shows the similarity matrix of pairwise Spearman correlations across labeled isotopomers.

Figure A1.S4— Association between nucleotide synthesis and redox metabolism.



Supplementary Figure 4

Scatterplots of average pathway expressions in TCGA colon cancer samples. A significantly positive association (regression p-value<0.0001) is seen in all three cases between NADPH pathway and glutathione, purine, or thymidine pathways.

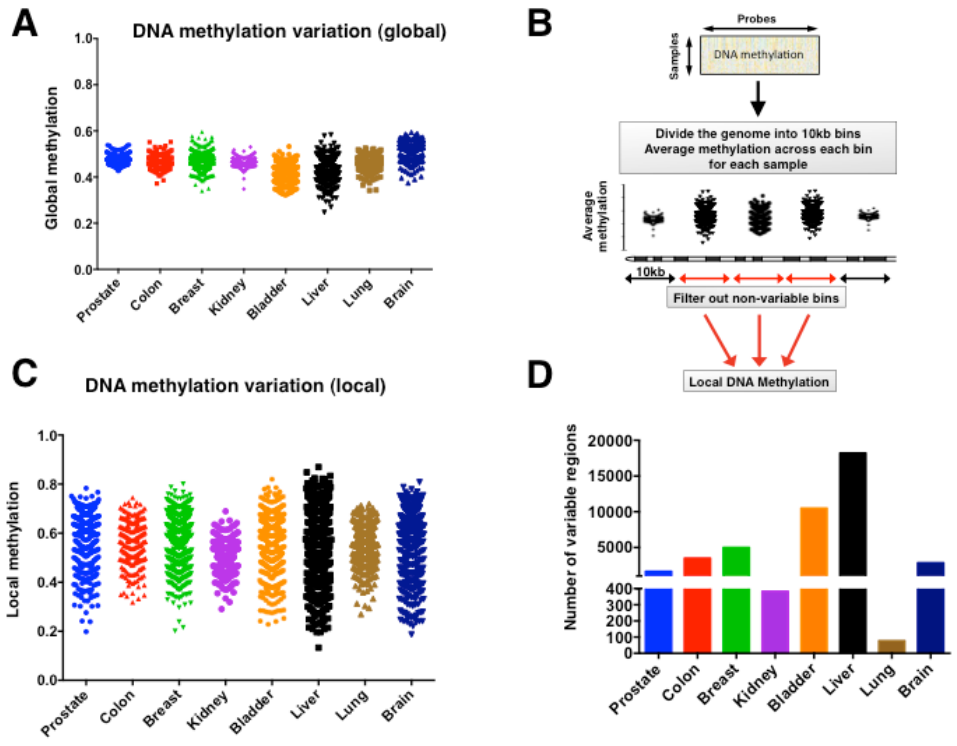
Table A1.S1— Calculated SGOC network fluxes across 8 cell lines.

	SW620	SW480	HCT8	HT29	HCT116	NCI-H508	SW48	SW948	MC-ave	MC-sd	MC-cv
Fshmt1+	7.79	3.44	1.41	1.61	2.96	6.56	1.64	1.21	2.99	0.26	0.09
Fshmt1-	7.19	2.37	0.32	0.70	2.07	5.35	0.59	0.25	2.18	0.21	0.10
Ftr_gly	0.48	0.26	0.58	0.34	0.32	0.28	0.48	0.66	0.33	0.06	0.18
Fx_ser+	0.55	0.10	0.00	0.21	0.31	0.28	0.03	0.09	0.33	0.11	0.33
Fphgdh	0.06	0.12	0.09	0.08	0.15	0.44	0.07	0.01	0.13	0.04	0.35
Fshmt2+	0.46	0.06	0.00	0.17	0.27	0.24	0.04	0.05	0.34	0.11	0.32
Fshmt2-	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
Fgcs	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.00	0.00
Fx_methf+	0.46	0.06	0.01	0.17	0.26	0.23	0.04	0.04	0.33	0.11	0.32
Fx_methf-	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.00
Fdilmethf	1.00	5.00	5.00	1.00	1.00	1.00	1.00	10.00	1.00	0.00	0.00
Fthddn	0.54	0.99	0.36	0.10	0.18	0.91	0.57	0.70	0.28	0.03	0.11
Fadedn	0.76	0.59	0.54	0.57	0.75	0.81	0.58	0.75	0.91	0.02	0.02
Fx-gly	0.51	0.12	0.05	0.27	0.35	0.28	0.07	0.05	0.38	0.11	0.28

Estimated fluxes and associated fitting errors are listed. All fluxes are fitted or calculated based on flux balance (exceptions are gcs, shmt2-, x_methf- and dil_methf that were fixed). Results from a Monte-Carlo simulation using parameters that reflect HCT116 cells are also shown (MC-ave: average flux from 500 iterations; MC-sd: standard deviation; MC-cv: coefficient of variation).

APPENDIX 2: SUPPLEMENTARY INFORMATION FOR CHAPTER 3.

Figure A2.S1—Pan-cancer analysis of global and local DNA methylation variation.



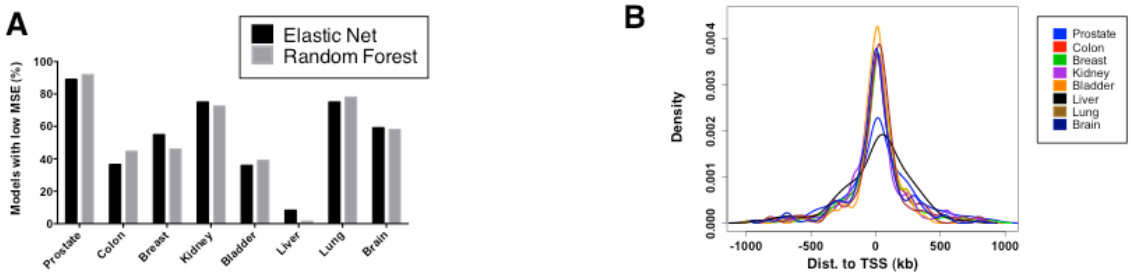
A) Variations in global DNA methylation are shown as measured by averaging the genome-wide value per sample. Values range between 0 and 1, with 1 indicating maximum methylation. Each point represents a unique tumor.

B) Schematic summarizing the approach used for dividing the genome into 10 kb intervals and calculating local DNA methylations.

C) Variations in local DNA methylation. Each point represents a unique sample, and local DNA methylation is calculated as the average value across all selected 10 kb regions.

D) Total number of 10 kb bins across the genome with variable DNA methylation (sd > 0.2) is shown for each cancer type.

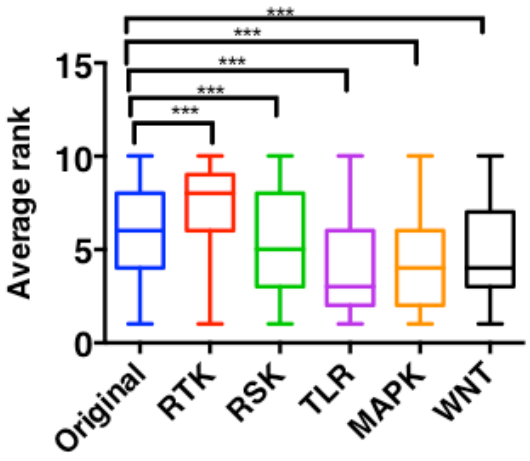
Figure A2.S2— Assessing models of local DNA methylation.



A) The y-axis shows the fraction of regions where DNA methylation was predicted with MSE smaller than 0.04 using the integrative models in each cancer.

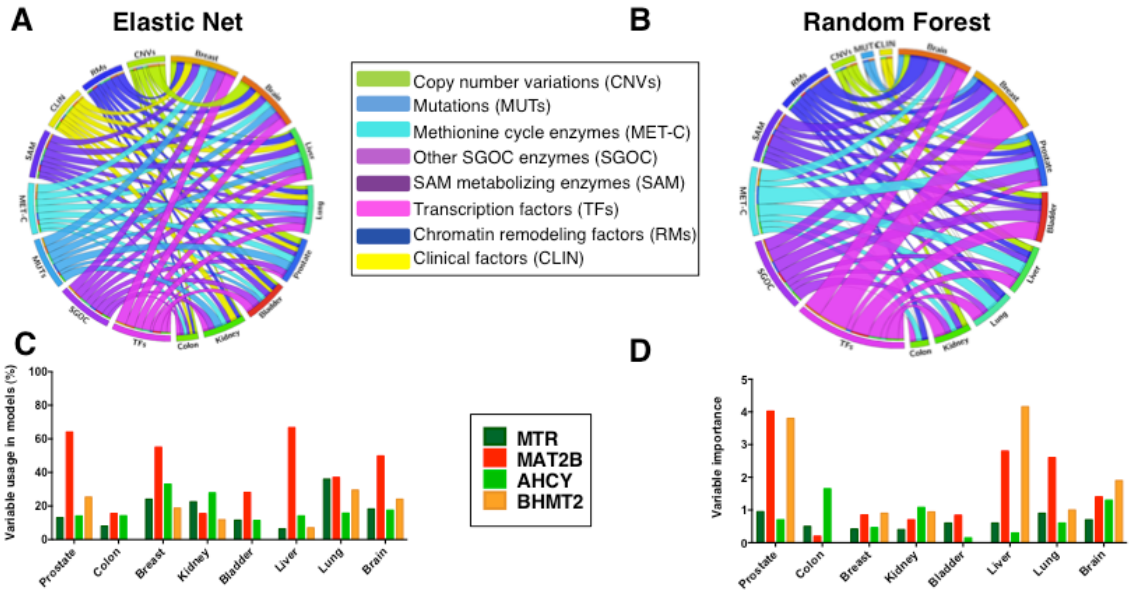
B) Density plots resulting from positional annotation of regions where DNA methylation was most predictable (smallest MSEs) by the integrative models. Distance to nearest gene's transcription start site (TSS) is shown on the x-axis in kilobases.

Figure A2.S3— Comparison of my gene expression variables with popular gene families.



Comparison of original gene expression variables with 5 popular gene sets was considered: Receptor tyrosine kinases (RTK), Receptor serine kinases (RSK), Toll like receptors (TLR), MAPK signaling (MAPK) and WNT signaling (WNT) families. The y-axis shows the average rank of each gene expression category based on average variable importance score across all Random Forest models of local DNA methylation in brain cancer (Error bars show the minimum and maximum value in each group). Significance of p-values associated with the Mann-Whitney test between the ranks across all models is shown (***: <0.0001; see Methods).

Figure A2.S4— Results of modeling local DNA methylation.



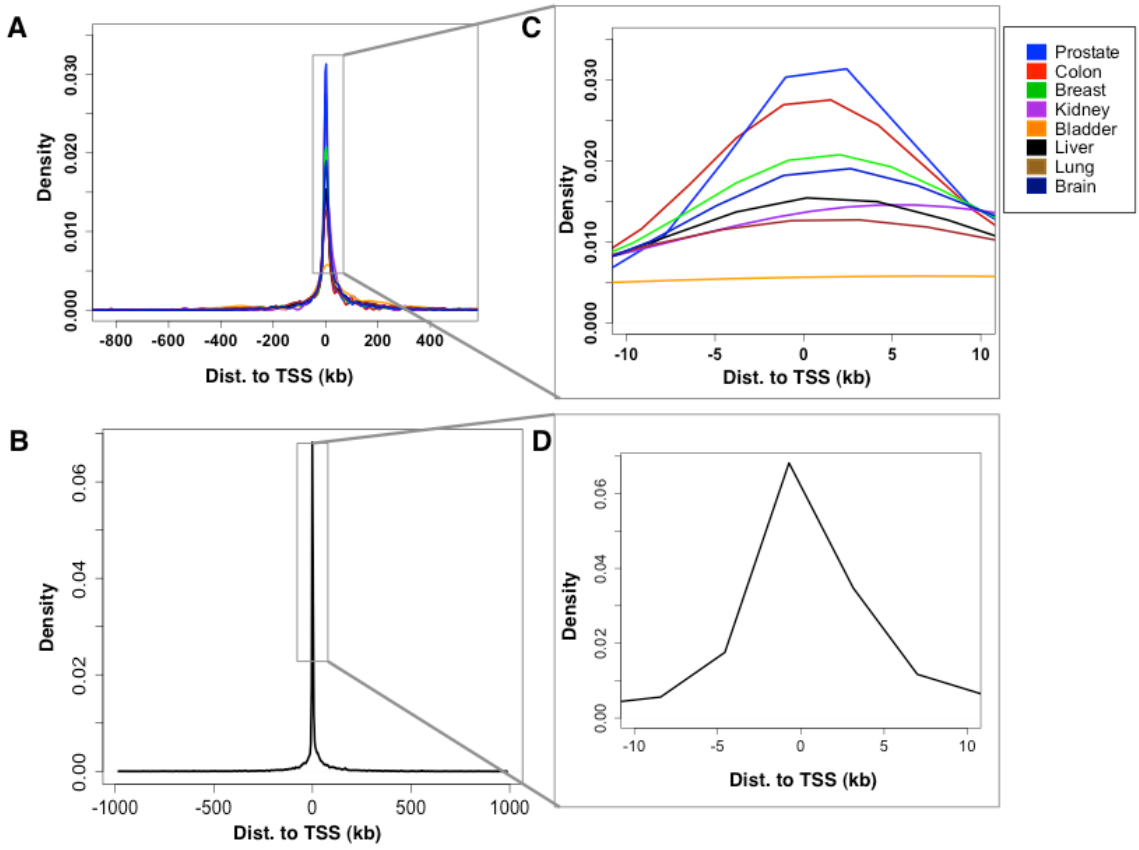
A) Contributions of variable classes to local DNA methylation are shown according to Elastic Net average variable usage (see Methods). The width of a given ribbon represents the relative value for the contribution of the corresponding variable class in the corresponding cancer type, with thicker ribbons showing higher relative contributions.

B) Contributions of variable classes to local DNA methylation are shown according to Random Forest average variable importance (see Methods).

C) The relative contributions of met cycle variables to local DNA methylation were calculated according to the Elastic Net integrative models with MSE<0.04. The y-axis shows the fraction of the 10 kb regions wherein each variable was selected for prediction of DNA methylation (variable usage).

D) The relative contributions of met cycle variables to local DNA methylation were calculated according to the Random Forest integrative models with MSE<0.04. The y-axis shows average variable importance score across all models.

Figure A2.S5— Annotation and evaluation of metabolically regulated regions.



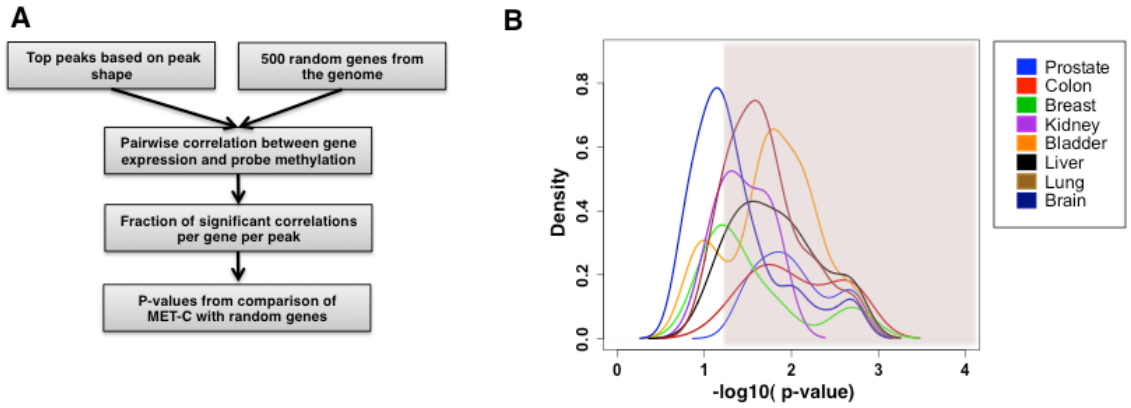
A) Density plots resulting from positional annotation of peaks identified in each cancer type by the genome-scanning algorithm described in Figure 3.3 are depicted. Distance to nearest gene's TSS is shown on the x-axis in kilobases.

B) Density plots of the distribution around nearest gene's TSS for 10000 randomly selected probes along the Illumina Infinium HumanMethylation 450K BeadChip arrays are shown.

C) Zoomed-in view from part "A" to visualize the distribution of peaks immediately surrounding the TSS region.

D) Zoomed-in view from part “B” to visualize the distribution of probes immediately surrounding the TSS region.

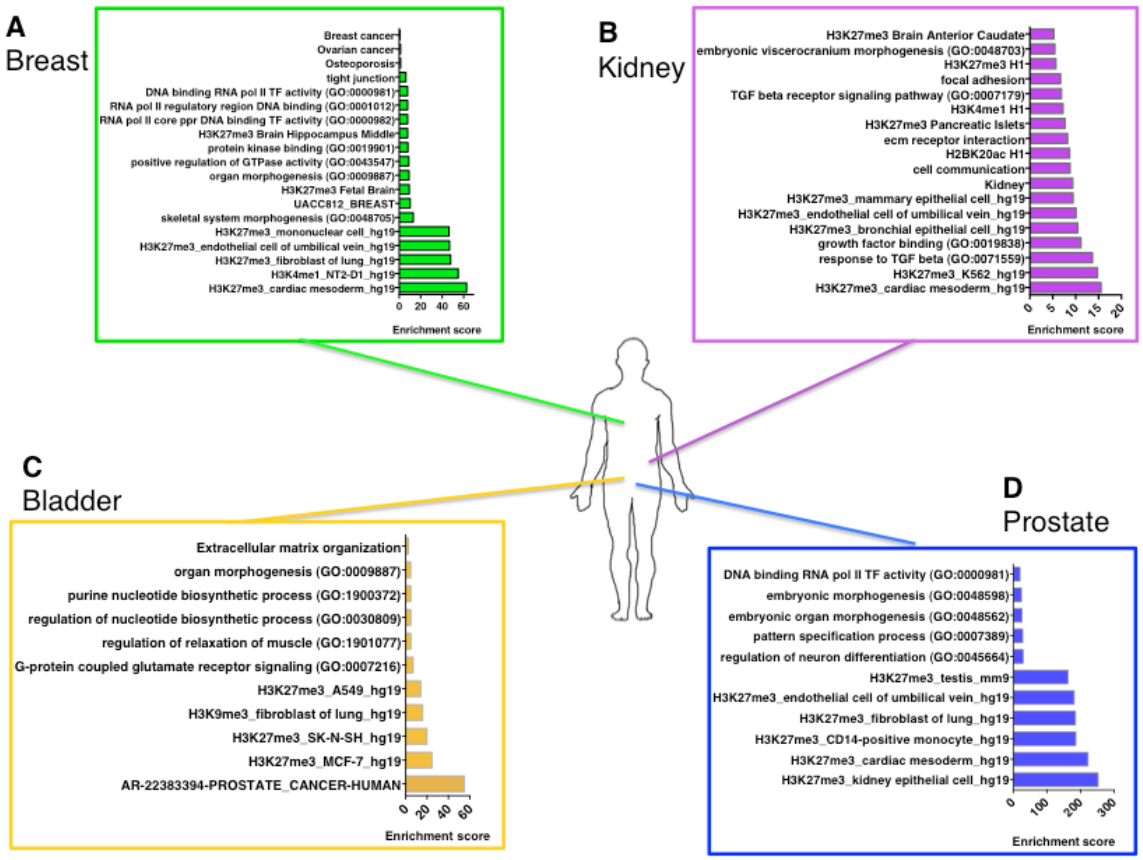
Figure A2.S6— Test of specificity of the metabolically regulated regions for correlation with met cycle expression.



A) Diagram describing the method used for testing specificity of correlation peaks for the met cycle genes vs. random genes. A p-value is calculated for each peak by comparison to 500 random genes (see Methods).

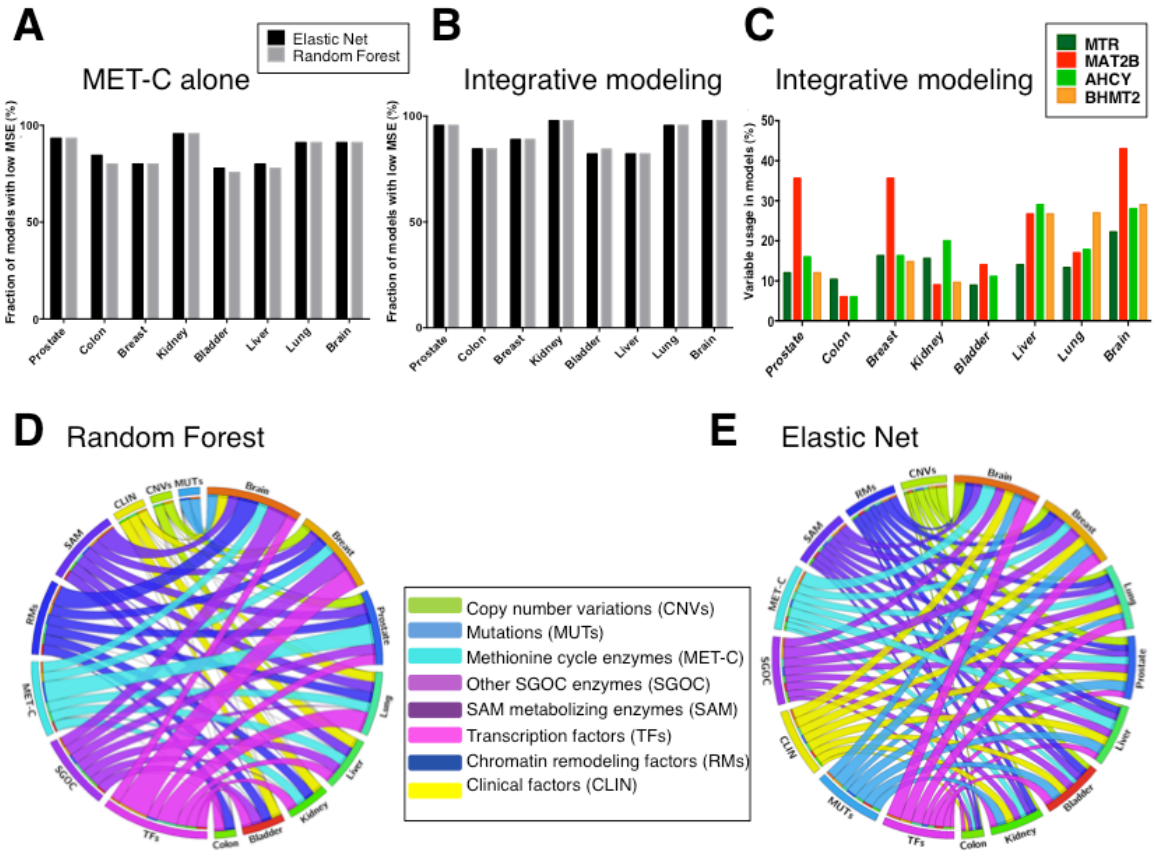
B) Density plot of the distribution of randomization p-values for all peaks. Shaded area shows significant p-values (<0.05) indicating peaks that were specifically and non-randomly correlated with the met cycle genes.

Figure A2.S7— Functional annotation of metabolically regulated regions of the epigenome.



A-D) Pathway enrichment analyses results in cancers of breast, kidney, bladder, and prostate. Functional annotation analyses were performed on lists of genes located within peaks of correlation between met cycle and DNA methylation in corresponding cancers.

Figure A2.S8— Modeling DNA methylation at cancer gene promoters and gene bodies.



A) Fraction of cancer genes that were predictable by met cycle genes alone with test set prediction error (MSE) of 0.01 or smaller.

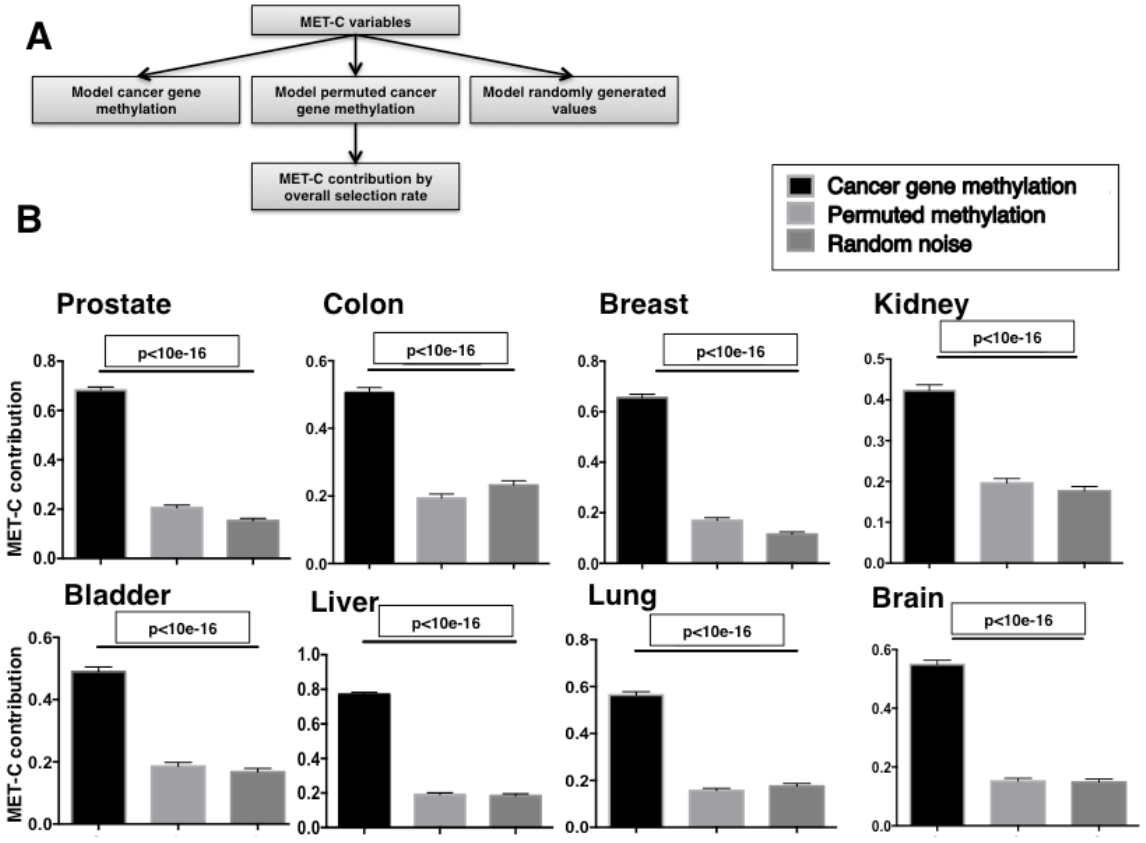
B) Fraction of genes that were predictable by the integrative models with test set prediction error (MSE) of 0.01 or smaller.

C) Fraction of Elastic Net models in which the met cycle variables was selected by the integrative approach in each cancer type.

D) Contribution of variable classes to cancer gene DNA methylation according to Random Forest average variable importance of each class (see Methods). The width of a given ribbon represents the relative value for the contribution of the corresponding variable class in the corresponding cancer type, with thicker ribbons showing higher relative contributions.

E) Contribution of variable classes to cancer gene DNA methylation according to Elastic Net average variable usage of each class (see Methods).

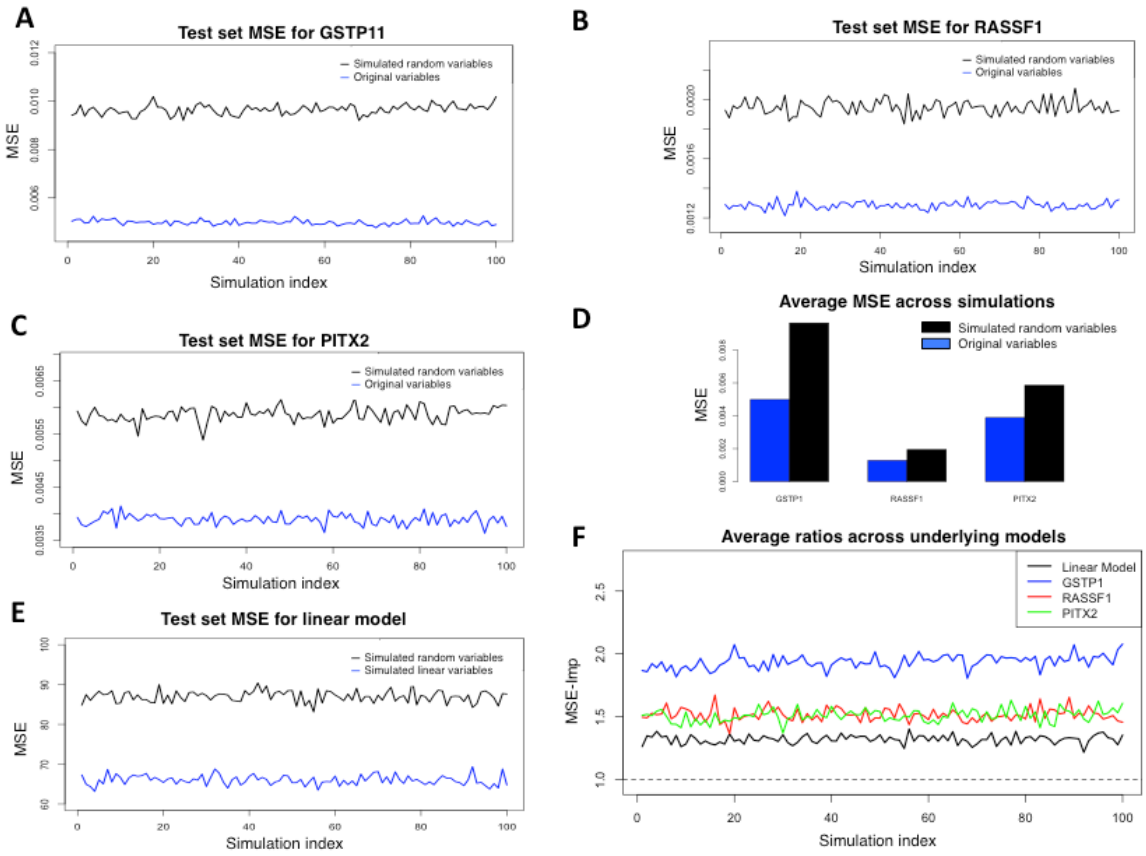
Figure A2.S9— Evaluation of modeling performance using randomized responses.



A) Diagram summarizing the approach used for testing the reliability of models by comparing cancer gene methylation values with randomized responses.

B) Average contribution of met cycle variables to prediction of cancer gene methylation vs. permutated methylation values and randomly generated numbers (see Methods). The y-axis shows the fraction of Elastic Net models wherein met cycle variables were selected. Kolmogorov-Smirnov non-parametric p-values were calculated between the variable usage values obtained using the original methylations vs. permutated or random responses separately. (Significant p-values (<10e-16) were also obtained by comparing the Random Forest variable importance scores across the models in all cases (not shown). Error bars show the standard error of mean (SEM) in each category).

Figure A2.S10— Evaluation of modeling performance using randomized predictors.



A) Comparison of glutathione S-transferase pi 1 (GSTP1) methylation prediction error by the original variables vs. random simulated variable sets of the same dimensions in prostate cancer (see Methods).

B) Comparison of RAS association domain family member-1 (RASSF1) methylation prediction error by the original variables vs. random simulated variable sets of the same dimensions in prostate cancer.

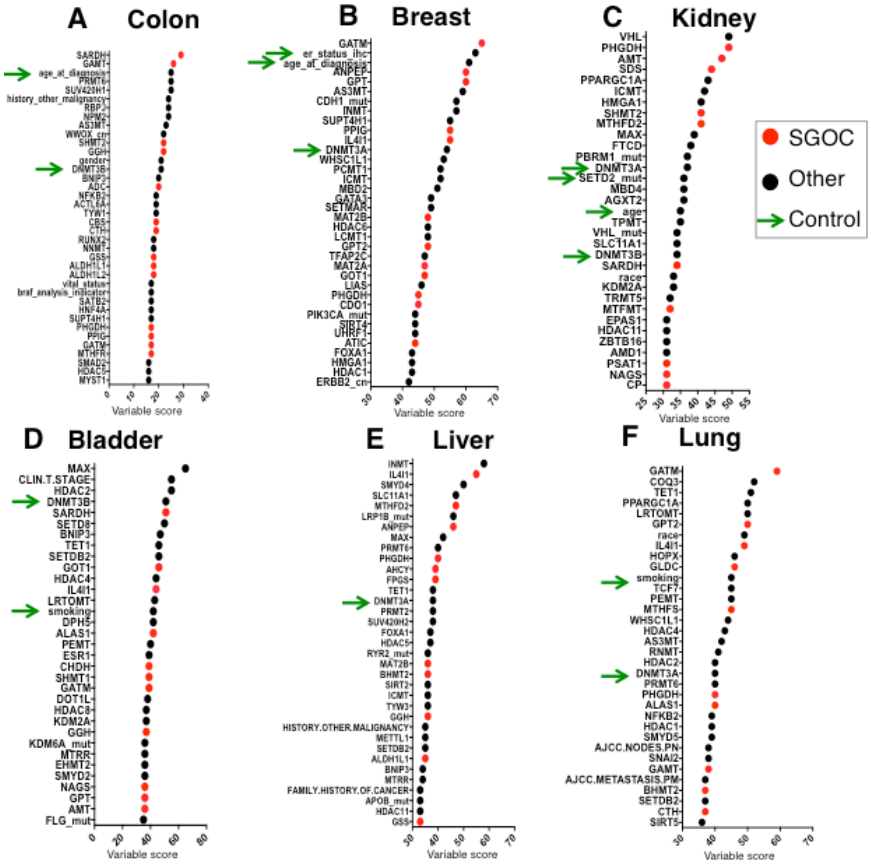
C) Comparison of paired-like homeodomain transcription factor 2 (PITX2) methylation prediction error by the original variables vs. random simulated variable sets of the same dimensions in prostate cancer.

D) Average MSE across 100 simulations of random predictors is shown for each of the responses.

E) Comparison of prediction errors for a simulated response by variables linearly related to the response vs. random variable set of the same dimensions.

F) Improvement of predictions by original variables vs. random variable (MSE-Imp= MSE-rand / MSE-orig) is plotted for the three example responses from my original dataset and also a simulated linearly-related dataset (see Methods). (MSE-rand= average MSE calculated using the randomly simulated variables, MSE-orig= average MSE calculated using the original variables)

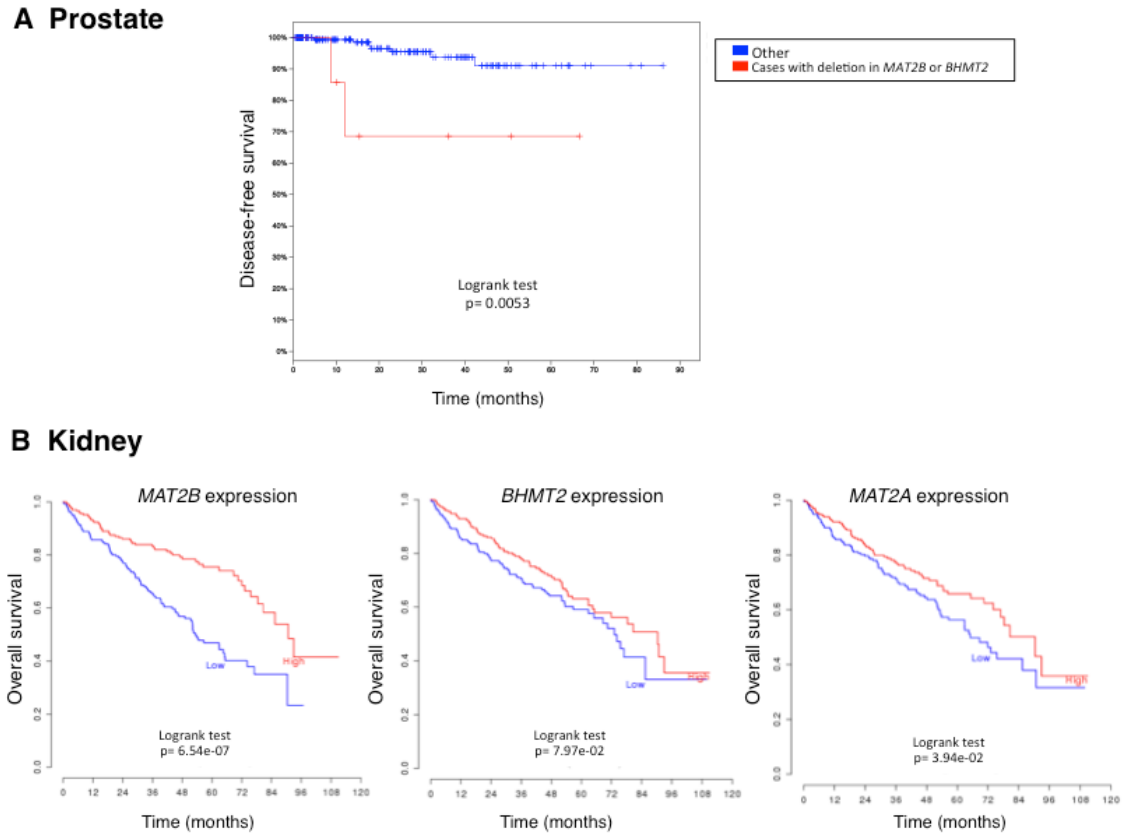
Figure A2.S11— Summary of predictive modeling of DNA methylation at cancer gene loci.



A-F) Variables that were most predictive of cancer gene methylation on average (top 15%) are listed and ranked in order of increasing contribution (variable score= percent variable usage by Elastic Net averaged across all models of cancer gene body and promoter methylation). Variables in the serine, glycine, one-carbon (SGOC) network (including the met cycle genes and other SGOC genes) are shown in red and all other variables are shown in black. Green arrows point to previously published factors associated with variations in DNA methylation in each cancer type (positive controls). (Variable names: official gene symbols are used to show gene expression variables

(including “Methionine cycle enzymes”, “Other SGOC enzymes”, “Transcription Factors”, “Chromatin Remodelers”, and “ SAM-metabolizing enzymes”), while “_mut” and “_cn” suffixes following gene symbols denote “Mutations” and “Copy Number Variations”, respectively. For “Clinical factors”, variable names match the descriptors used in the TCGA clinical data files).

Figure A2.S12— Independent analyses of survival in TCGA cases by cBioPortal and PRECOG.



A) Comparison of disease-free survival between patients with deep deletions in *MAT2B* or *BHMT2* genes and other patients in the TCGA prostate cancer cohort. The plot and the log-rank test p-value were adopted from the cBioPortal.

B) Comparison of overall survival between TCGA kidney cancer patient groups exhibiting high expression vs. low-expression of the met cycle genes. Plots and log-rank test p-values were adopted from Prediction of Clinical Outcome from Genomic profiles (PRECOG). (MAT2B= methionine-adenosyltransferase 2B, MAT2A=

methionine-adenosyltransferase 2A, BHMT2= betaine-homocysteine S-methyltransferase 2)

Note A2.S1—Variation in DNA methylation.

I considered an analysis of a large set of DNA methylation arrays from the TCGA that were collected and processed according to a standardized procedure that results in an estimate of the relative amount of DNA methylation at each oligonucleotide probe (the beta-value). This value ranges from 0 to 1 with 1 indicating that each allele is completely methylated(Cancer Genome Atlas, 2012). Arrays were used over bisulfite sequencing because of the higher availability of these data in a standardized format allowing for an integrative analysis. I considered several tumor types with large sample sizes where both RNA-seq and DNA methylation data was available on each tumor (breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC), brain lower grade glioma (LGG), bladder urothelial carcinoma (BLCA), kidney renal clear cell carcinoma (KIRC), and prostate adenocarcinoma (PRAD).

Upon analysis of global DNA methylation levels (average per tumor), I observed that variation in global methylation across tumors from the same cancer type is higher than the between-cancer-type variation (between-cancer-type sum of squares (SS)= 44%, within-cancer-type SS= 56%) (Figure A2.S1A). This differs from what is typically seen in normal tissues where between-tissue type variability in DNA methylation exceeds within-tissue type variability by an order of magnitude(Lokk et

al., 2014; Ziller et al., 2013). Thus, my results confirm increased inter-individual variation in DNA methylation among tumors from the same tissue of origin, consistent with methylation hypervariability in cancer(Hansen et al., 2011). It is important to note however, that due to the nature of the current TCGA data (one sample per tumor), I were unable to further parse this inter-individual variation to distinguish between variations caused by differences between individual patients *vs*. differences between clonal populations of cells within a given tumor.

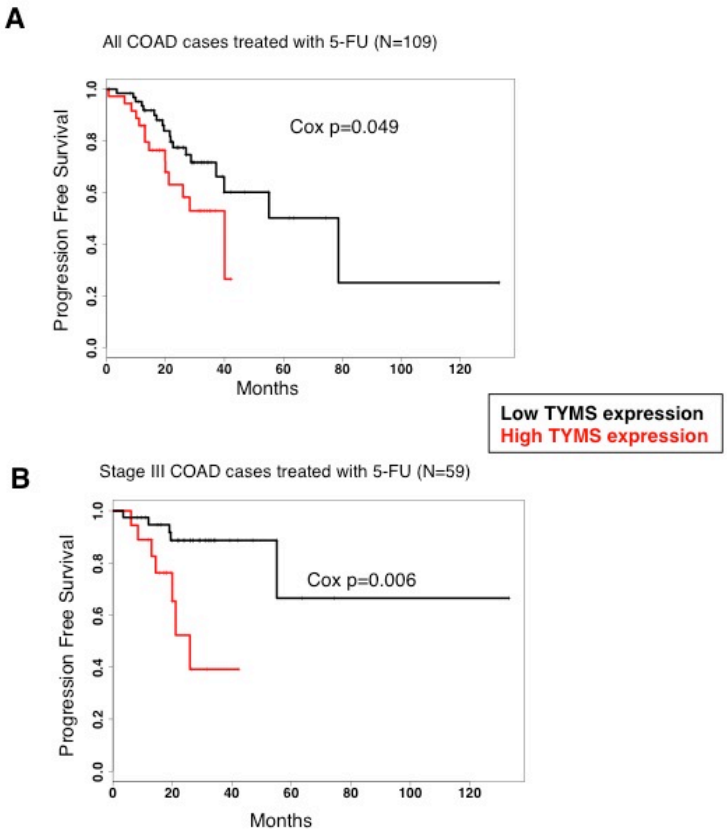
Since the biological function of DNA methylation occurs at specific regions of genomic DNA, I considered a local analysis of DNA methylation. I partitioned the genome into 10 kilobase (kb) regions and calculated average methylation in each region (Figure A2.S1B; Methods). Notably, a previous study showed that DNA methylation at genomic regions with high inter-individual variation is more likely to be associated with expression of nearby genes, suggesting that variable regions are enriched for functionally active DNA methylation(Gutierrez-Arcelus et al., 2013). I therefore focused only on regions with standard deviation (sd) of 0.2 or higher for the subsequent integrative analyses (Figure A2.S1C). The number of such regions differed substantially among cancer types, with liver and bladder cancers exhibiting the largest number of variable DNA methylation regions (Figure A2.S1D).

References A2.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.
Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., et al. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife *2*, e00523.
Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nature genetics *43*, 768-775.
Lokk, K., Modhukur, V., Rajashekar, B., Martens, K., Magi, R., Kolde, R., Koltsina, M., Nilsson, T.K., Vilo, J., Salumets, A., et al. (2014). DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. Genome biology *15*, r54.
Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477-481.

APPENDIX 3: SUPPLEMENTARY INFORMATION FOR CHAPTER 5.

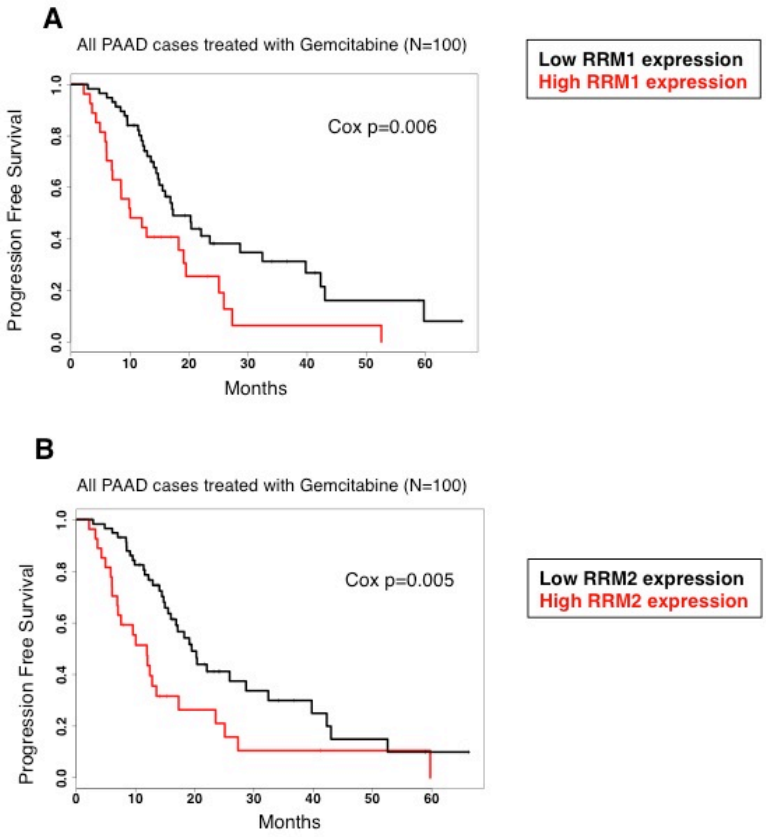
Figure A3.S1—Relationship between target enzyme expression and response to 5-FU in colon cancer.



A) Kaplan-Meier plot compares progression free survival in high-TYMS expression vs. low-TYMS expression subgroups of TCGA COAD patients.

B) Kaplan-Meier plot compares progression free survival in high-TYMS expression vs. low-TYMS expression subgroups of stage III TCGA COAD patients.

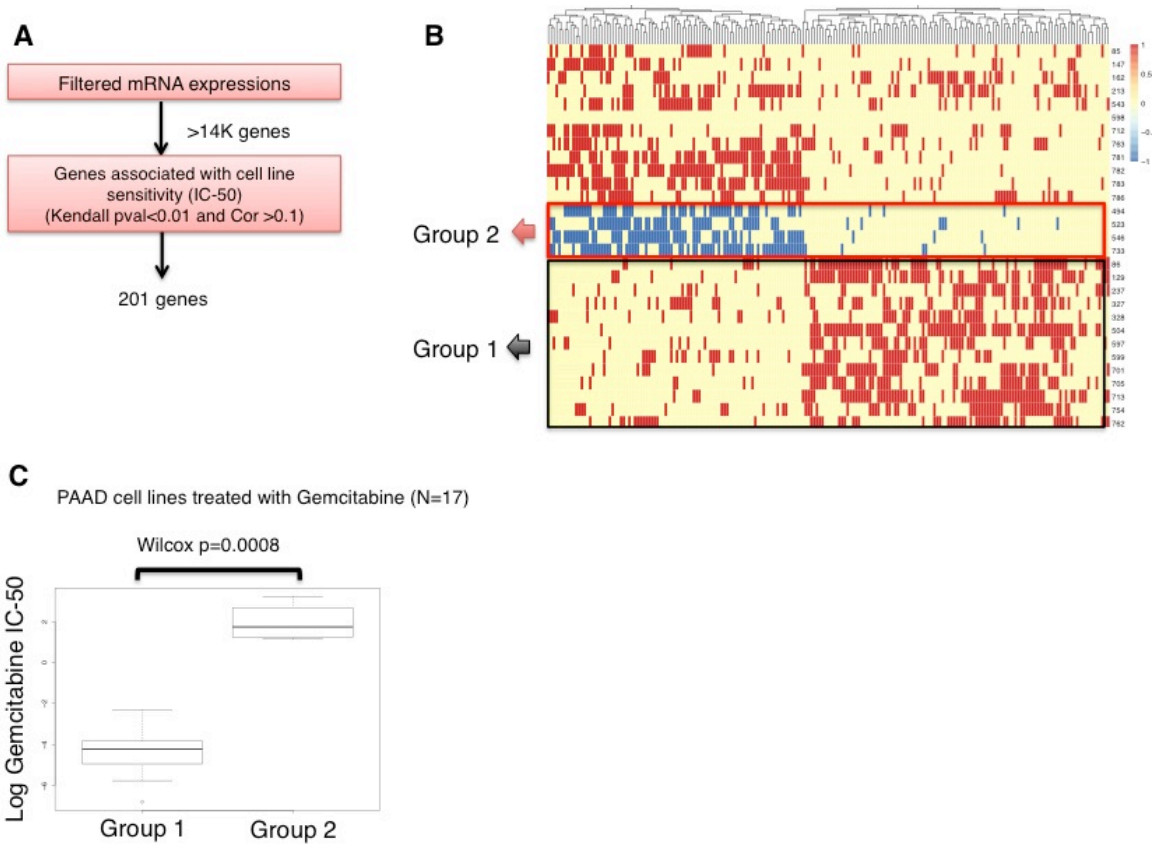
Figure A3.S2—Relationship between target enzyme expression and response to Gemcitabine in pancreatic cancer.



A) Kaplan-Meier plot compares progression free survival in high-RRM1 expression vs. low-RRM1 expression subgroups of TCGA PAAD patients.

B) Kaplan-Meier plot compares progression free survival in high-RRM2 expression vs. low-RRM2 expression subgroups of TCGA PAAD patients.

Figure A3.S3— Identifying gene expression signatures of sensitivity to Gemcitabine in pancreatic cancer cell lines.



A) Schematic of the step-wise filtering used for gene selection in pancreatic cancer (COSMIC PAAD).

B) Hierarchical clustering heatmap of the discretized gene favorability scores. Columns represent genes and rows represent individuals. Favorable scores are shown by the color red (F=1), unfavorable by blue (F= -1), and neutral by yellow (F=0) (see Methods).

C) Box-plots comparing the resistance to Gemcitabine (log IC-50 values) between the two cell line subgroups identified in part B (error bars show the range of the data points in each group).